Spring 5-2012

Assessing the Functional Accuracy and Quality of Volunteered Geographic Information: A Comparison of Open Street Map and Navteq Road Datasets

Brett William Hode
University of Southern Mississippi

The University of Southern Mississippi

ASSESSING THE FUNCTIONAL ACCURACY AND QUALITY OF VOLUNTEERED GEOGRAPHIC INFORMATION: A COMPARISON OF OPEN STREET MAP AND NAVTEQ ROAD DATASETS

by

Brett William Hode

A Thesis
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Master of Arts

Approved:

Dean of the Graduate School

May 2012

ABSTRACT

ASSESSING THE FUNCTIONAL ACCURACY AND QUALITY OF

VOLUNTEERED GEOGRAPHIC INFORMATION:

A COMPARISON OF OPEN STREET MAP AND NAVTEQ ROAD DATASETS

by Brett William Hode

May 2012

This thesis provides a detailed analysis of the functional data accuracy between Open Street Map data and Navteq road data. The analysis revealed that the average accuracy level ranged from 87.3% to 94.9% for buffer distances between 2 and 20 meters. Analyses were also performed to determine the predictability and spatial distributions of accuracy levels. The results showed that there is no statistical relationship between population density, education levels, or poverty levels when compared to the accuracy levels of OSM data. Further, no clearly discernible patterns in the spatial distribution of accuracy values for OSM data were found. The overall conclusion is that the expected accuracy of OSM data is comparable to that of commercially available solutions, and the only limitation on the use of this data is its use in routing critical services due to a lack of sufficient attribute information on many of the roadways in the OSM dataset.

ACKNOWLEDGMENTS

I would like to thank my thesis committee for their time and input into the research and thesis preparation. I would also like to thank my employer, Naval Research Laboratory, for supporting this research effort by providing data and computational resources, as well as the time and flexibility in my work schedule required to complete this Master's degree program.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure

LIST OF TABLES

Table

LIST OF ACRONYMS

ESRI	Environmental Systems Research Institute
GB	Giga-Byte
GPS	Global Positioning System
ISO	International Organization for Standardisation
NSSDA	National Standard for Spatial Data Accuracy
OGC	Open Geospatial Consortium
OSM	Open Street Map
TIGER	Topologically Integrated Geographic Encoding and Referencing system
UN	United Nations
UNITAR	United Nations Institute for Training and Research
VGI	Volunteered Geographic Information
XML	eXtensible Markup Language

DEFINITIONS

* All definitions in this section were taken from Kresse and Fadaie (2004)

Buffer - Geometric object that contains all direct positions whose distance from a specified geometric object is less than or equal to a given distance (ISO39 2003)

Positional Accuracy (Absolute) - Closeness of coordinate value to the true or accepted value in a specified reference system

Relative Positional Accuracy - Closeness of coordinate difference value to the true or accepted value in a specified reference system (ISO48 2002)

Quality - Totality of characteristics of a product that bear on its ability to satisfy stated and implied needs (ISO34 2002)

Reference Data -Data accepted as representing the universe of discourse, to be used as reference for direct external quality evaluation methods (ISO46 2001)

Accuracy - Closeness of agreement between a test result and the accepted reference value (ISO45 2002, ISO 48 2002)

Dataset - identifiable collection of data

Geographic Information System - Information system dealing with information concerning phenomena associated with location relative to the Earth. (ISO34 2002)

Precision - Measure of the repeatability of a set of measurements (ISO48 2002)

Completeness - Presence and absence of features, their attributes and relationships

Logical Consistency - Degree of adherence to logical rules of data structure, attribution and relationships.

Temporal accuracy - Accuracy of the temporal attributes and temporal relationships of features.

Thematic accuracy - Accuracy of quantitative attributes and the correctness of non-quantitative attributes, as well as the classification of features and their relationships.

CHAPTER I

INTRODUCTION

Geographic information systems (GIS) professionals have been significant adopters of both open source software and open data products. As a result open source GIS tools have become full featured applications that rival ESRI and other commercial solutions. This type of development has occurred through collaborative efforts in software development a trend that has taken place since the beginning of the home-computing industry and termed open-source software in 1998 with the founding of the Open Source Initiative (Open Source Initiative, 2010). Recently, a new trend has developed encouraging the open sharing and collection of data. This geographic data sharing practice has been aptly named Volunteered Geographic Information (VGI) (Goodchild, 2007; Hall et al., 2010). VGI data can typically be described as open data, or data that can be contributed to, reproduced, modified, and redistributed without legal barriers (Science Commons, 2008). Typically, this type of data is generated through collaborative efforts in spatial data collection, commonly referred to as community or participatory mapping (Perkins, 2007; Goodchild, 2007), in which data is voluntarily contributed to the larger effort to collectively enhance the dataset.

Open spatial data is becoming extremely important in modern geography as new technologies and reduced prices in commodity technologies have enabled VGI mapping data to be collected over large spatial extents. Chief among the changes in technology are the reduced price of digital storage media, global positioning system (GPS) devices, and an increase in the availability of high resolution imagery. In the past, data was collected selectively because of high collection and storage costs. However, the cost to store data

has dropped significantly, since April 1995 the cost per gigabyte of storage has dropped from 625.00 U.S. Dollars to less than $0.08 in modern storage devices (Alts.net, 2008). GPS devices have seen similar trends in cost reductions making them a common device in many individuals' everyday life (Hakley & Weber, 2008; Goodchild, 2007). Finally, despite efforts to enhance publically available data, governments simply do not collect the data that is needed for many projects leading amateurs to create the data that is needed from available resources and collaborative efforts (Goodchild, 2007; Wood, 2005).

Accurate road data is critical to many areas of research, government management, and humanitarian efforts. Most commonly used commercial solutions such as Navteq or Tele Atlas, or government data products such as TIGER are used to fulfill this requirement. Ironically, the areas that are most in need of such data often have little to no available data due to either financial or technical constraints. This research was undertaken to address the feasibility of using a volunteered geographic information product, Open Street Map, in professional GIS projects in the United States. Open Street Map (OSM) data is the focus of most VGI related efforts in the media and academia, particularly after it was used as the primary dataset for road features during the 2010 Haiti earthquake disaster relief effort by many non-governmental organizations (NGOs), government agencies, and military groups (Unitar, 2012; OpenStreetMap, 2010e, National Oceanic and Atmospheric Administration, 2010). OSM has been adopted by the United Nations Institute for Training and Research (UNITAR) as the basis for its community mapping projects (along with other technologies) in areas such as the Horn of Africa, Sudan, Southeast Asia, to address natural disasters, human displacement (e.g.

refugees) among other issues through the production of professional GIS datasets and mapping products (Unitar, 2012).

Open Street Map Dataset Description

Open Street Map (OSM) is a vector road dataset that is free to edit and use. Developing the dataset involved contributions from almost 500,000 users and has collected almost 2.75 billion GPS track points (Open Street Map, 2010d). The result of this collection effort is a global road dataset that contains over 113 million "ways" (road feature segments). The data is eXtensible Markup Language (XML) encoded based on an ontological schema for describing features established by the Open Street Map Foundation but also allows for the creation of custom schema's for specific purposes. OSM created the United States portion of its dataset by converting the U.S. Census Department's TIGER line data into the OSM XML schema and users then contributed additional data to, in theory, improve the quality and accuracy of the OSM data (OpenStreetMap, 2011). TIGER is one of the most used datasets for the U.S. because it is a free road data network that is relatively complete for the entire United States making it comparable to Navteq or OSM. OSM was the most common VGI data product in professional and academic literature found for this thesis, and because of its large contributor base, this dataset was selected as the basis for evaluation against a dataset with a well-respected level of accuracy, Navteq (Hakley, 2010; Ludwig et al., 2011; Zielstra & Zipf, 2010; Feilner, 2009; Science Daily, 2007; Linux Pro Magazine, 2010).

Navteq Dataset Description

Navteq is a global road, geocoding, and points of interest data product that is widely used in commercial, non-profit, and government GIS projects. Navteq provides road and traffic information in the commercially available Garmin car navigation systems (e.g. Garmin Nuvi GPS systems) (Garmin, 2012; Privat, 2011; Menga, 2007; Microsoft, 2012). The use of Navteq data in many different GIS applications establishes it as a defacto industry standard for critical applications. For the purposes of this thesis this dataset will be the independent "true-earth" representation of the real world road features in place of in situ data which is impractical to attempt to collect for such a large area.

Research Justification

This research was necessary to evaluate the suitability of using OSM data in professional research and GIS projects. There are few existing evaluations of OSM data accuracy that were found and none evaluated its accuracy in the United States (Hakley, 2010; Ludwig et al., 2011; Zielstra & Zipf, 2010). Studies were typically conducted in Europe and were limited to much smaller areas with fewer features than what was used in this thesis (Hakley, 2008; Zielstra & Zipf, 2010). Further, evaluating the United States is a critical step because the U.S. is one of the most car-centric cultures in the world due to a lack of a highly effective mass transit system (Pentland, 2008). The utility of OSM has been seen repeatedly in disaster recovery efforts. Any level of data accuracy and data quality is a benefit in areas that have little or no access to government data or the financial means to acquire commercial data. The essential case study in how VGI data can be more viable than traditional data sources was seen in the use of OSM data in Haiti, where thousands of contributors used satellite imagery and various other sources to

effectively map an entire country in a few days. The data was then exported into various formats and web services to enable search and rescue personnel, military personnel, and the general public to access the data (OpenStreetMap, 2012).

Figure 1. Port-au-Prince, Haiti before the January 12, 2010 earthquake. The image shows that only major roads were mapped and many appear as broken line segments (Maron, 2010).

Figure 2. Port-au-Prince, Haiti after the January 12, 2010 earthquake. The figure shows large sections of the city filled with road data collected after the earthquake between January 10, 2010 and January 14, 2010 (Maron, 2010).

The use of OSM data in such critical roles represents a shift in mentality and functionality requirements being pushed by the web 2.0 environments that are ubiquitous in many people's daily life. The intent of the research is to explore whether that desire for interactive data is well placed and whether it exhibits any patterns that can be used to predict spatial distribution of acceptable levels of accuracy.

Research Questions

The purpose of this research is to examine the suitability of OSM data in terms of positional accuracy and data quality as described in ISO 19113:2002 and 19114. The objectives are to determine what factors may contribute to the quality of the data such as population density, and evaluate the accuracy of the data sets with regard to identified factors. To accomplish this goal three hypotheses and corresponding research questions were explored based on the body of literature dealing with quality assessment of spatial datasets.

1) Open Street Map data is comparable to commercial data solutions in terms of positional accuracy.
    a. What is the typical level of data accuracy of Open Street Map data?
    b. Does the level of functional data accuracy vary significantly between regions, or other areas with specific characteristics?
2) Population density is a direct contributor to the accuracy of VGI data because it is entirely user generated.
    a. Does Population distribution and density affect the functional accuracy of Open Street Map data?

b. Is there a revealing pattern at regional, state, or local level regarding data accuracy? Is accuracy correlated to other characteristics such as poverty levels, or average education?

3) OSM data will exhibit an acceptable level of data quality compared to Navteq data products.

a. How comparable to a commercial data product is Open Street Map data in terms of completeness and semantic quality?

b. Does the ability for users to readily edit OSM data result in a lower accuracy in terms of semantic quality?

CHAPTER II

REVIEW OF RELATED LITERATURE:

There is a variety of literature that was reviewed in order to develop the methodology that will address the research questions listed in chapter 1. The literature can be broken into several categories: (1) general VGI use, (2) accuracy assessment techniques, and (3) existing OSM-centric research. Each will be discussed in depth in this chapter.

VGI Background Information

Volunteered Geographic Information (VGI) is a relatively new topic in GIS research and almost all academic literature related to VGI has been published in the past 5 years. There has been a similar increase in the appearance of VGI efforts in trade journals and magazines (Feilner, 2009; Science Daily, 2007; Linux Pro Magazine, 2010; ESRI, 2010). In addition to these sources, there has been an increase in the occurrence of OSM data in professional cartographic products. Many websites are utilizing OSM data as the vector road product such as the United Nations UNITAR division who adopted OSM as a primary data source for cartographic products (Unitar, 2012). VGI and OSM-related articles have become increasingly common in both non-peer review and peer reviewed periodicals indicating that VGI efforts are a topic of growing interest and use in GIS projects (Elwood, 2009; Goodchild, 2007; Perkins, 2007). The increased use of VGI is expected because VGI projects have the potential to obtain information that is not normally collected (Goodchild, 2007; Wood, 2005). VGI techniques are now being used extensively for data collection outside academia because of reductions in the price of storage media, GPS devices, and other technology (Goodchild, 2007; Hakley & Weber,

2008). Existing literature also suggest that the increasing availability of high-bandwidth internet connections (Goodchild, 2007), Web 2.0 interfaces (Elwood, 2009; Hall et al., 2010), and a reduction in the amount of publicly available data for end users to exploit (Goodchild, 2007; National Geospatial Intelligence Agency, 2005; Wood, 2005) may have contributed to the rapid rise of these types of social mapping endeavors.

In community mapping two trends exist. First these projects tend to fill a requirement for data that is not provided effectively by government agencies or when commercial solutions are not cost effective. Second, most projects avoid professional standards in favor of ease of development and use to encourage participation. (Perkins, 2007; Wood, 2005). The first trend has the potential to significantly benefit geographers by providing additional data that can be used in research efforts. The second trend tends to create problems with data consumption in the professional world as commercial tools, like ESRI's ArcGIS software, are often not compatible with the data products produced through Participatory or VGI efforts. The positive aspects of VGI have even been adopted by major commercial geospatial data providers such as Google and Navteq who now have sites where users can report issues with the produced data, make edits, and there is a change review process which allows appropriate changes to be made (Navteq, 2011; Google, 2011).

Quality Assessment Techniques

The International Standards Organization (ISO) defines five types of data quality in standard ISO 19113:2002 outlined in table 1 (Kresse & Fadaie, 2004).

Table 1

ISO 19113:2002 Data Quality Metrics

Term	Definition
Completeness	Presence and absence of features, their attributes and relationships.
Logical Consistency	Degree of adherence to logical rules of a data structure, attribution and relationships (data structure can be conceptual, logical or physical).
Positional Accuracy	Accuracy of the position of features
Temporal Accuracy	Accuracy of the temporal attributes and temporal relationships of features.
Thematic Accuracy	Accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships.

Note: Definitions taken from Kresse and Fadaie (2004)

Each of these metric must be considered when assessing dataset quality and accuracy. However, completeness, logical consistency, temporal accuracy, and thematic accuracy are very difficult to assess and very little literature was found that establish ways of measuring these metrics. Temporal accuracy is often easy to determine but is difficult to establish what constitutes an acceptable level of temporal accuracy. The body of academic literature is primarily focused on methods for positional accuracy assessment to examine the accuracy of geographic features. Positional Accuracy can be described in two ways: First, as precise accuracy or the exact difference in position of a feature in a dataset relative to the real-world position of the feature. Or second, in terms of functional accuracy, also referred to as relative accuracy, which is the difference in position between

a feature in a dataset and the same feature in a dataset with a high level of accuracy that can be accepted as a good representation of the real-world features of interest (Van Niel & McVicar, 2002).

Precise accuracy assessment techniques vary by country but in the United States is typically described by the National Standard for Spatial Data Accuracy (NSSDA) created by the USGS as a way to measure the radial error at any given point along a vector feature (Federal Geographic Data Committee, 1998). The NSSDA method calculates the difference of the real values and the independent test case, uses these values to calculate the RMSE value, and multiplies it by a correction factor. This technique is ideally suited for point data and requires a minimum of 20 clearly defined points with exact in-situ measurements of the real world position to be statistically accurate. Figure 3 below is an example form that is used by The State of Minnesota to perform road feature accuracy assessments that utilizes the NSSDA method to determine the positional accuracy that can be expected for a dataset at a 95% confidence level. Van Niel and McVicar (2002) concluded that the most error-prone location in a point and line network is the intersections of lines, and they showed that along a line points tended to be more accurate when assessing them using the NSSDA method. The NSSDA method may be a commonly used precise accuracy assessment technique in the United States. However, it is not well suited for non-normal data distributions or large area studies due to its reliance on in-situ data collection (Zandbergen, 2008). While a standard for data accuracy assessment the NSSDA method is not well suited for the large areas covered in this study.

The seminal literature in the field of functional accuracy assessment is Goodchild and Hunter (1997) which describes a methodology to compare vector features by creating a series of buffers around a reference dataset that represents the real-world position of the features being examined. The technique does not rely on in situ data, which makes it possible to work with large areas where detailed in situ information is difficult to obtain. The test dataset with an unknown positional accuracy is then buffered with a one meter buffer and the overlap is calculated as a simple ratio of the overlap of each buffer to determine the distance between the reference line and the test dataset equivalent line generating a percentage of the test dataset that falls within each increment in distance of the buffer around the reference dataset.

Figure 4. Horizontal accuracy statistic worksheet.

A Point number	B Point description	C x (inde-pendent)	D x (test)	E diff in x	F (diff in x)^2	G y (inde-pendent)	H y (test)	I diff in y	J (diff in y)^2	K (diff in x)^2 + (diff in y)^2
									sum	
									average	
									RMSE$_r$	
									NSSDA	

Column	Title	Content
A	Point number	Designator of test point
B	Point description	Description of test point
C	x (independent)	x coordinate of point from independent data set
D	x (test)	x coordinate of point from test data set
E	diff in x	x (independent) - x (test)
F	(diff in x)^2	Squared difference in x = (x (independent) - x (test))^2
G	y (independent)	y coordinate of point from independent data set
H	y (test)	y coordinate of point from test data set
I	diff in y	y (independent) - y (test)
J	(diff in y)^2	Squared difference in y = (y (independent) - y (test))^2
K	(diff in x)^2 + (diff in y)^2	Squared difference in x plus squared difference in y = (error radius)^2
	sum	$\sum$ [(diff in x)^2 + (diff in y)^2]
	average	sum / number of points
	RMSE$_r$	Root Mean Square Error (radial) = average$^{1/2}$
	NSSDA	National Standard for Spatial Data Accuracy statistic = 1.7308 * RMSE$_r$

Figure 3. State of Minnesota NSSDA Positional Accuracy Calculation Worksheet. This is provided as an example of how the NSSDA method is utilized in a concise format. This figure was taken directly from The State of Minnesota Positional Accuracy Handbook.

OSM-Centric Research

This section will systematically cover the academic and non-peer reviewed articles that directly relate to the use, accuracy, and quality of Open Street Map data. Despite the increase in the use of OSM data and popularity of other VGI datasets as a research topic there is a limited body of academic literature and most of it has been published in the past five years. Because this thesis is focused on answering questions related to the quality, completeness, and data utility of OSM data this section is dedicated to examining the existing research on this specific dataset.

The body of OSM research is highly European-centric to date; every accuracy or quality assessment found performed an assessment of an European country or using subsets of an European country. Due to the limited areas that have been analyzed by the existing body of literature it is necessary to begin analyzing OSM data in other areas to determine its true suitability as a data product. A review of the current literature on OSM accuracy analysis techniques was conducted to determine the best methodology to use for this research. Perhaps the most critical study of OSM data previously performed was Hakley (2010) in which OSM data was compared to the United Kingdom Ordinance Survey dataset. Hakley (2010) found that OSM data was approximately 85% similar to the Ordinance Survey dataset. This study utilized the Goodchild and Hunter (1997) method discussed previously to determine the degree of overlap between the two datasets. This study was cited by all other papers performing comparisons of OSM data to established datasets representing seminal literature in the area of OSM data accuracy assessment (Ludwig et al., 2011; Zielstra & Zipf, 2010). The methods used by Hakley (2010) were used as the basis for conducting this study. Since the methodology of Hakley

(2010) has been tested in repeated academic literature, only the location of the study and the post-accuracy assessment statistical analyses changed between studies the underlying methods were the same.

Ludwig et al. (2011) attempted to analyze OSM data against Navteq data products. In their study the researchers used feature matching techniques based on five fields mutual to the Navteq and OSM datasets (road type, road name, direction of travel, speed limits, and pedestrian access) for detailed analyses. They conducted this study for Germany. They utilized buffers around the Navteq dataset at five, ten, and thirty meters to segment the OSM dataset into smaller pieces to match the encoding they observed in the Navteq dataset. They found that to accurately match features they had to exclude many pieces of modern road networks stating only that "incompatible categories will be discarded" (Ludwig et al., 2011) without detailing how much or the types of data being discarded. This study was highly reliant on data completeness for all attributes in both datasets and returned useful results showing that between 44% and 82% of OSM road features were found within five meters of the Navteq counterpart depending on the area. Further, the study showed that near 100% of OSM features were within thirty meters of the Navteq feature, and it showed that OSM dataset was significantly less complete in terms of attribute accuracy, often due to missing information. OSM attribute information completeness was shown to vary between 79.8% complete to as low as 50.8% in rural areas.

The methodologies used in the Ludwig et al. (2011) study cannot be applied to a study in the United States, due to several differences in the data. First, the available OSM dataset for the U.S. does not record the street speed limits, pedestrian access, and a quick

assessment shows that the U.S. OSM dataset has very low completeness for values other than road type and road name. Based on a visual inspection of road data tables for the U.S. OSM counties often had less than 2% of roads marked with their lane count or directionality, and as much as 40% of features lacked street names. Due to the limited attribute information it was determined that feature matching Navteq with the United States OSM data would be impossible. It should be noted that quite a few of the road features that lacked names were service roads, on/off ramps, and similar features which may not have an actual road name so this may not indicate a lack of quality instead the figures may legitimately may not require names. Navteq data was consistently complete in terms of attribute information; however, in the United States it appears that the OSM dataset is not consistent in recording attribute information. This lack of data consistency in each dataset makes attribute-based feature matching algorithms impractical because only a few features could accurately be matched between datasets in each county.

Zielstra and Zipf (2010) performed a similar analysis of OSM compared to Tele Atlas, another commercial solution similar to Navteq. The paper compared all roads in Germany between the two datasets. This paper found that OSM contained up to 30% less total road length compared to Tele Atlas, and that only 50-85% of OSM data fell within a ten meter buffer in their tested areas. These figures agree with the Ludwig et al. (2011) accuracy assessment for Germany using Navteq features. The Tele Atlas dataset was buffered and then how much of roads fell within a ten meter buffer was measured. Given the similarity in methodology between all existing accuracy studies conducted with OSM data, the Goodchild and Hunter (1997) method will be used to perform the accuracy assessment for this thesis.

CHAPTER III

METHODOLOGY

Study Area

The study area consists of the southeastern United States including: Florida, Georgia, South Carolina, North Carolina, Alabama, Mississippi, Louisiana, Tennessee, and Arkansas. These states comprise a total of 755 counties. Wilson County Tennessee was excluded due to error that could not be corrected in the Navteq dataset that prevented it from being processed into an usable format. This leaves a total sample size of 754 counties (approximately 24% of the 3136 United States).

Dataset Description

The following datasets were obtained either through the manufacturer or through contracts available to the author as a federal employee. All datasets were used within the restrictions of their license terms.

1. Navteq 2011 Road Layer: This dataset contains a complete road map of the United States and is considered a de-facto standard for road data in the GIS industry. Navteq road datasets are commonly used in commercial GPS devices and professional GIS projects; as such it will be used as the baseline for acceptable data for the purposes of this thesis.
2. Open Street Map 2011: obtained August 2011 for the entire United States in 51 shape files from the Open Street Map foundation website. The shape files were chosen to make it easier to utilize the OSM data in

the data analysis and to allow the use of ESRI Application Programming Interfaces to perform data preparation.

3. U.S. Census Bureau 2010 county level demographic data set. Although all values were recorded in the dataset, for processing the fields used were population density per square mile, percent population in poverty, and percent population with a bachelor's degree.
4. U. S. Census Bureau 2010 TIGER county outlines shape files by state. These files were used to clip the Open Street Map and Navteq datasets by a standard polygon for each county-level dataset that would coincide with available demographic data.

Data Preparation Application

As previously stated the goal of this study is to analyze the functional accuracy of Open Street Map Data compared with Navteq datasets. The primary datasets were originally obtained in two formats. The Navteq dataset was obtained from Homeland Security in a compressed ESRI format. This Smart Data Compression (SDC) format is a highly compressed shapefile format designed to allow shape files, traditionally limited to 2 Gigabyte in size, to contain significantly more data. The primary challenge in working with SDC formatted files is the increased time to access features due to decompression operations and the total file sizes which prevents certain analyses from being run in a reasonable amount of time. The Navteq datasets required several days to process and resulted in over 100 Gigabytes of data for the Navteq dataset. This dataset is divided into many layers reflecting road features, points of interest, and various other categories of data. Only the road features were selected for use in this research. The Navteq road

dataset is divided into several layers which had to be evaluated to determine what best reflects a single continuous layer for the continental United States. The separate layers reflected different recommended levels of detail for each level of zoom that would typically be used in a GIS; the finest level of detail was retained to ensure that the most detailed and most spatially accurate data was retained. The finest level of detail should have the fewest generalizations in the line geometry reflecting a better "real-world" version of the road features. The SDC formatted Navteq data covered the entire United States, due to time restrictions only the southeastern United States were selected for this thesis. To extract the Navteq data the SDC file was broken into county-level non-compressed shape files. County-level files enable a granularity of data that is consistent with available demographic data and are smaller than the file size limit of the shape file format.

The Open Street Map dataset is developed in a XML file format and is also exported as a shape file. OSM shape files were downloaded for each state as a shapefile and then broken into county-level files using an ESRI arcpy script. Both datasets were clipped using the same county-outlines obtained from the U. S. Census Bureau 2010 census TIGER cartographic boundary files. Before each dataset was clipped they were all projected into WGS 84 Datum to match the TIGER cartographic boundary files. Once clipped, each county-level file was projected into North American Equidistant Conic Projection to preserve the length of all road features when analyses were performed. A new field was added to all shape files and it was populated with the length of each road feature. Next, the Navteq county-files were buffered using a series of ten incremental buffers starting at two meters and increasing by two meters each time up to twenty

meters. Once all of the buffers were created the buffered files were used to clip each of the OSM dataset files to produce new OSM files. The feature lengths are automatically recalculated when the files are created. All the files were then summed to produce a comma separated values file that contains the feature lengths for each county, as well as the percentages of the feature lengths that fell within each buffer when compared with the feature lengths of the original dataset.

The data preparation application then reads the shape files and generates the following Comma Separated Values (CSV) file entries for each county. Each row represents a column in the CSV file. A complete description of the recorded fields is available in Appendix A. Once the CSV file was created an additional set of values were appended to the table taken from the U.S. Census department demographic data for each county. The census department records a variety of data for each county, data recorded for each county as part of the 2010 census were appended to the CSV files. The U.S. Census data fields are recorded in Appendix B. Figure 4 shows the process flow of the data preparation methodologies. Once the data is in the CSV file it is possible to perform analyses to answer the hypotheses posed in this thesis. All statistical analyses were performed using SPSS PASW 18. Most of the hypotheses were able to be answered using basic descriptive statistics, including mean accuracy level, minimum, and maximum accuracy levels for each buffer distance, and frequency distributions.

All application development and calculations were performed using Python 2.7 and the ESRI ArcGIS python application programming interface (API). The processing was performed on two workstation systems with two quad-core processors and 12GB of memory each. The systems were configured to run six county files concurrently using

multiprocessing techniques and averaged 14 minutes per county to process the data. The total computing time was a little less than 200 hours and generated 500GB of data in approximately 950,000 files.

To analyze how closely the OSM dataset's positional accuracy matches the accuracy of the Navteq dataset the mean positional accuracy levels were computed for each buffered distance. The functional accuracy of Open Street Map is easily represented by the mean positional accuracy for each state and for each buffered distance using the entire southeastern region. A regional accuracy level is established by the mean accuracy level for each buffered distance. Although accuracy is computed for each of the ten buffer distances the most important are six, eight, ten, and fourteen meters because they are close representations of typical road widths based on the Florida roadway design manual (State of Florida, 2012). To determine if the datasets are similar in terms of their data accuracy a Kruskal-Wallis (K-W) test was performed using each state as a grouping variable to determine if there was significant variance between states. A K-W test was selected because the data was found to be non-parametric; this test reveals if there are significant differences in variance between each of the nine states. This test was only performed at state level due to the lack of clear divisions at county level to compare smaller regions. The K-W tests revealed that considering all states simultaneously they are only similar within the two, sixteen, eighteen, and twenty meter buffered distances. All other buffer distances showed some variation between states. A 2-independent samples (K-W) analysis comparing each state to each other state was performed to determine which states were similar, this test is intended to reveal if there are one or more states that are distinctly dissimilar from the others.

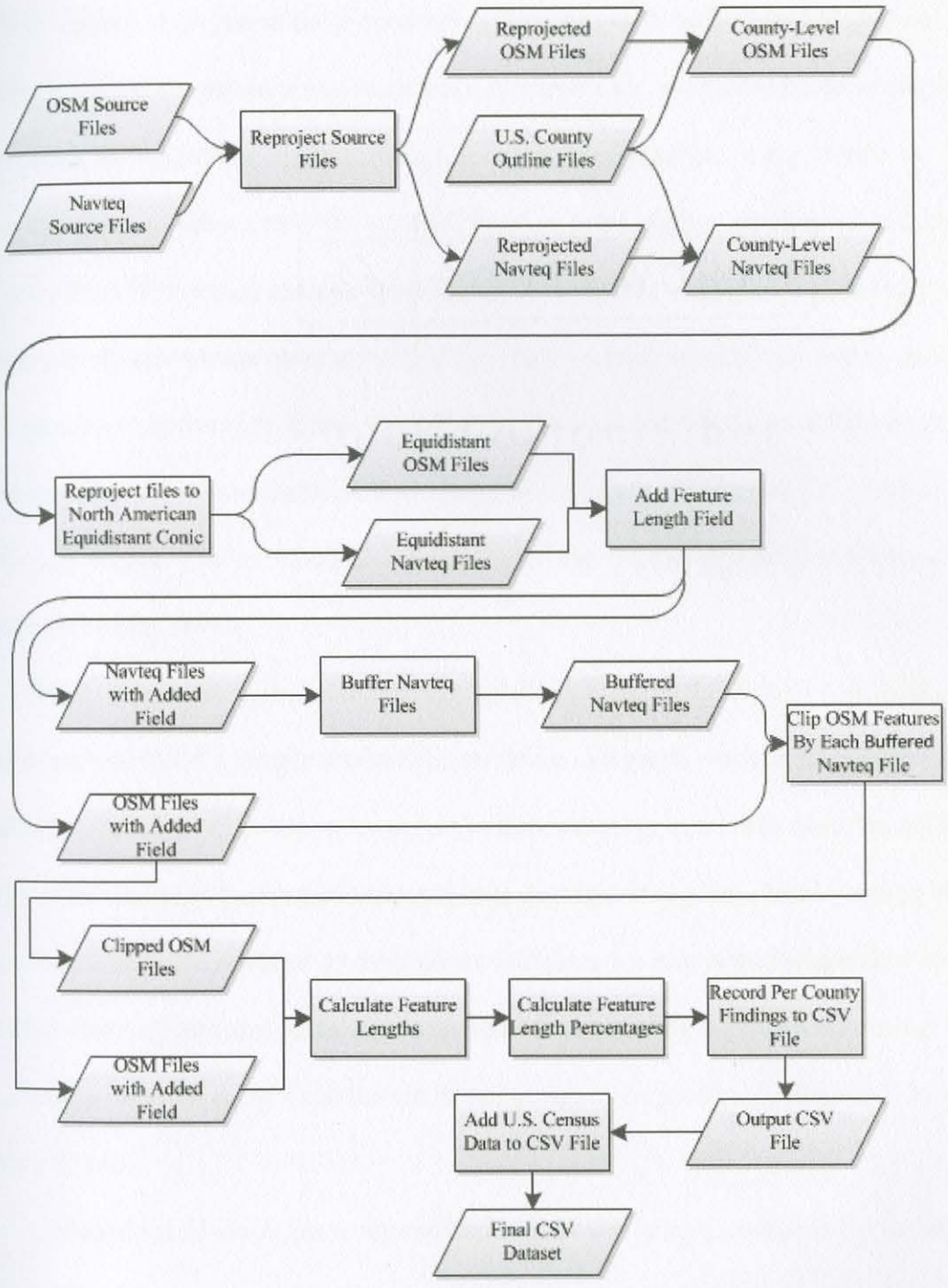

Figure 4. Data Processing Workflow Diagram. This figure is a step by step representation of the process that was carried out by the ESRI ArcGIS Scripts that were used to process the data prior to analysis.

A visual analysis was performed by creating a map in ArcGIS 10 that was color-coded by county to represent the percentage accuracy at each buffer level. Based on existing literature the minimum average accuracy levels for any buffer distance should be between 80- 85% (Hakley, 2010; Ludwig, 2011). If the average accuracy is 80% or higher it is deemed functionally acceptable based on prior studies. An average accuracy of greater than 90% would indicate that OSM is functionally comparable to the Navteq dataset. Ideally an average accuracy of 95% or higher would be achieved within the 20 meter maximum buffer size. If this level of functional accuracy is found at 20 meters or less buffer distance it is appropriate to describe OSM data as functionally equivalent to the Navteq dataset. The categorization displayed in the maps in figures 6-9 are based on these acceptability levels.

Next, to determine if various demographic data affects the accuracy of VGI datasets such as OSM a simple statistical correlation and graph was conducted the relationship between VGI data accuracy and various demographic data from the 2010 census. Analyses were performed on population density, educational levels, and poverty levels. These tests are intended to determine (1) if there are any relationships between VGI data accuracy and any of these demographic factors, and (2) if there are strong relationships between factors can the statistical accuracy be predicted reasonably by any of these factors.

To establish if OSM has a reasonable level of data quality compared to the Navteq dataset proved the most problematic to evaluate in any quantitative way, as such only completeness was analyzed using automated, quantitative techniques. All other metrics

used to define data quality established by the ISO were analyzed manually by reviewing a small subset (n=20) county files to record trends in the way data was recorded.

When examining data quality the International Organization for Standardisation sets out 5 metrics for geographic data quality: completeness, logical consistency, positional accuracy, temporal accuracy, and thematic accuracy. The methodology above was used to verify positional accuracy using the Goodchild and Hunter (1997) method. Next temporally these datasets are virtually identical with both being updated frequently, the only difference being that OSM is free to download so updates do not have to be negotiated as part of a service contract or purchased each time a researcher needs a new copy. Both datasets also match closely in logical consistency, as a minimum both datasets were found to record road type, and the feature geometry. Navteq has an additional 51 fields that are recorded whenever possible and with a great deal more consistency than the OSM dataset indicating that it has better feature attribute information completeness. To address data completeness an analysis was conducted by computing the ratio of OSM total road lengths to Navteq total road lengths to determine which dataset contained the most road data in meters. The result of this analysis provided a county-by-county basis for which dataset is more complete, a binary map was produced in Figure 19 illustrating which represents more total road length. The final quality metric: thematic accuracy, was not able to be analyzed due to differences in data encoding and no acceptable third party resources available to act as an intermediary to compare both datasets to.

CHAPTER IV

RESULTS

This section will outline the results of the statistical analyses described in the Methodology section. This section is organized to address each individual hypothesis and to draw conclusions about the hypothesis based on the results.

The positional accuracy assessment was conducted at ten buffer distances; however, three distances were more closely examined: six, eight, and fourteen meters are equivalent to typical two, three, and four lane road widths which constitute the vast majority of roads in the United States. Table 2 shows the accuracy levels by buffered distance for the southeastern United States. The highlighted entries show the most relevant functional accuracy assessments. The analyses have shown that 91% or better of all road features on average are within six meters of the equivalent Navteq road dataset. Further, 94% of all OSM road features can be expected to fall within fourteen meters of the equivalent Navteq road Dataset. OSM and Navteq datasets in limited areas can be functionally equivalent in terms of accuracy with several counties throughout the southeastern region scoring 99% or higher accuracy levels in one or more of the buffer distances.

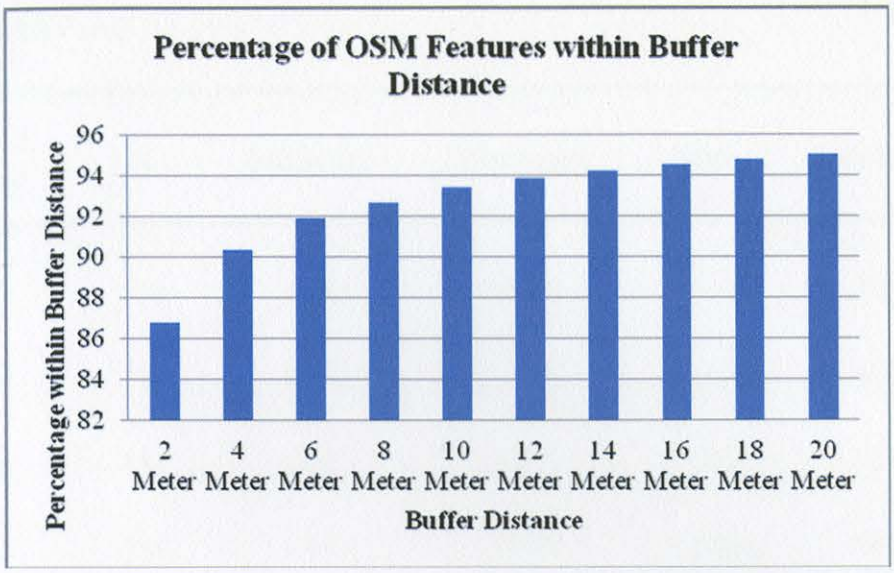

Figure 5. Average accuracy within each buffered distance analyzed. This chart shows the progression of accuracy levels for each buffered distance, it is clear that beyond 12 meters buffered distance the accuracy level apparently plateaus showing limited change in the amount of OSM road data within each increase in buffered distance.

Table 2

Mean Accuracy Levels per Buffered Distance

Buffered Distance	N	Minimum	Maximum	Mean	Std. Deviation
2 meters	749	.5470	.9970	.873265	.0601828
4 meters	754	.5753	.9988	.903161	.0539282
6 meters	754	.6238	.9989	.918842	.0506507
8 meters	753	.6490	.9991	.927850	.0480357
10 meters	754	.6577	.9991	.933772	.0457100
12 meters	754	.6642	.9991	.938353	.0438001

Table 2 (Continued).

Buffered Distance	N	Minimum	Maximum	Mean	Std. Deviation
14 meters	754	.6662	.9994	.941993	.0423486
16 meters	754	.6662	.9994	.945084	.0411886
18 meters	754	.6662	.9994	.947725	.0403255
20meters	754	.6662	.9994	.949894	.0396742
Valid N	748				
Average	.927993				

Note: This table shows the mean accuracy levels of Open Street Map data in decimal notation. The notable values are that 6, 8, and 14 meter buffered distances all exceed 90% functional accuracy when compared with the Navteq dataset. The 90% margin indicates that these datasets are functionally comparable to Navteq in terms of accuracy.

The second consideration in functional accuracy assessment was whether the accuracy varied by regions. The K-W tests showed that there was some variation in the accuracy from state to state at buffer distances of four to fourteen meters, buffer distances of two, sixteen, eighteen, and twenty meters were found to be statistically similar across all southeastern states (.878, .125, .183, and .167 significance respectively). The remaining buffer distances were not considered except for the six, eight, and fourteen meter buffer distances. These three were tested individually in a two independent sample K-W analysis. In each buffer distance case Alabama and Georgia were the only states that routinely showed a high degree of difference. Tables 3 through 5 below indicate the states that were similar or dissimilar for each buffer distance. The results showed that in buffer distances of 4-14 meters Georgia and Alabama are dissimilar from almost every

other dataset this shows that there is some state to state difference in accuracy levels but there is no discernible pattern as to why these two states are statistically different.

A visual inspection showed clear clustering of inaccurate areas primarily focused around the coastal areas of North Carolina, the mountainous areas in Tennessee, and the northwestern part of Mississippi. These three clusters are shown in Figure 6. However, as shown in Figure 7, most of these areas are only at unacceptable levels of accuracy at buffer distances of less than 6 meters. They remain the least accurate areas throughout the study. There are no definitive explanations for why these three clusters of unacceptable accuracy levels appear in these locations. However, one possible explanation for the reduced accuracy cluster in eastern Tennessee is that OSM data is partially compiled by digitizing satellite imagery, given that this area is a heavily forested and complex terrain that is prone to cloud cover making it difficult to digitize features accurately.

Table 3

6 Meter Buffer Distance Similarity (Kruskal-Wallis significance)

	Alabama	Arkansas	Florida	Georgia	Louisiana	Mississippi	North Carolina	South Carolina	Tennessee
Alabama		.000	.003	.336	.018	.000	.150	.171	.000
Arkansas	.000		.969	.004	.591	.847	.798	.321	.434
Florida	.003	.969		.012	.642	.967	.435	.290	.349
Georgia	.336	.004	.012		.062	.003	.285	.386	.000
Louisiana	.018	.591	.642	.062		.575	.769	.501	.163
Mississippi	.000	.847	.967	.003	.575		.615	.285	.284
North Carolina	.150	.798	.435	.285	.769	.615		.762	.186
South Carolina	.171	.321	.290	.386	.501	.285	.762		.044
Tennessee	.000	.434	.349	.000	.163	.284	.186	.044	

Table 4

8 Meter Buffer Distance Similarity (Kruskal-Wallis significance)

	Alabama	Arkansas	Florida	Georgia	Louisiana	Mississippi	North Carolina	South Carolina	Tennessee
Alabama		.001	.001	.237	.009	.000	.109	.098	.000
Arkansas	.001		.628	.009	.937	.969	.885	.439	.243
Florida	.001	.628		.003	.600	.717	.334	.183	.535
Georgia	.237	.009	.003		.043	.005	.275	.308	.000
Louisiana	.009	.937	.600	.043		.759	.631	.432	.234
Mississippi	.000	.969	.717	.005	.759		.633	.325	.241
North Carolina	.109	.885	.334	.275	.631	.633		.775	.162
South Carolina	.098	.439	.183	.308	.432	.325	.775		.053
Tennessee	.000	.243	.535	.000	.234	.241	.162	.053	

Table 5

14 Meter Buffer Distance Similarity (Kruskal-Wallis significance)

	Alabama	Arkansas	Florida	Georgia	Louisiana	Mississippi	North Carolina	South Carolina	Tennessee
Alabama		.036	.002	.163	.012	.039	.167	.096	.001
Arkansas	.036		.132	.363	.214	.969	.693	.856	.330
Florida	.002	.132		.015	.775	.143	.225	.218	.469
Georgia	.163	.363	.015		.069	.390	.575	.480	.020
Louisiana	.012	.214	.775	.069		.326	.366	.370	.689
Mississippi	.039	.969	.143	.390	.326		.939	.897	.276
North Carolina	.167	.693	.225	.575	.366	.939		.827	.493
South Carolina	.096	.556	.218	.480	.370	.897	.827		.497
Tennessee	.001	.330	.469	.020	.689	.276	.493	.497	

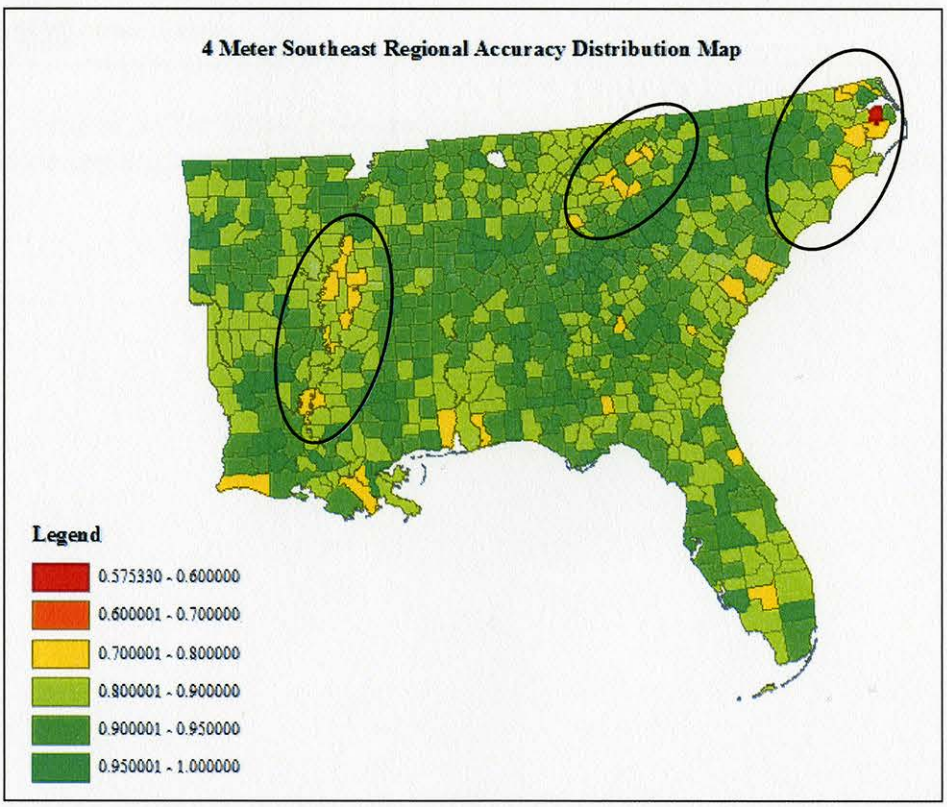

Figure 6. 4 Meter buffer distance accuracy distribution map. The map shows that there are areas of unacceptable data accuracy indicated by bounding circles.

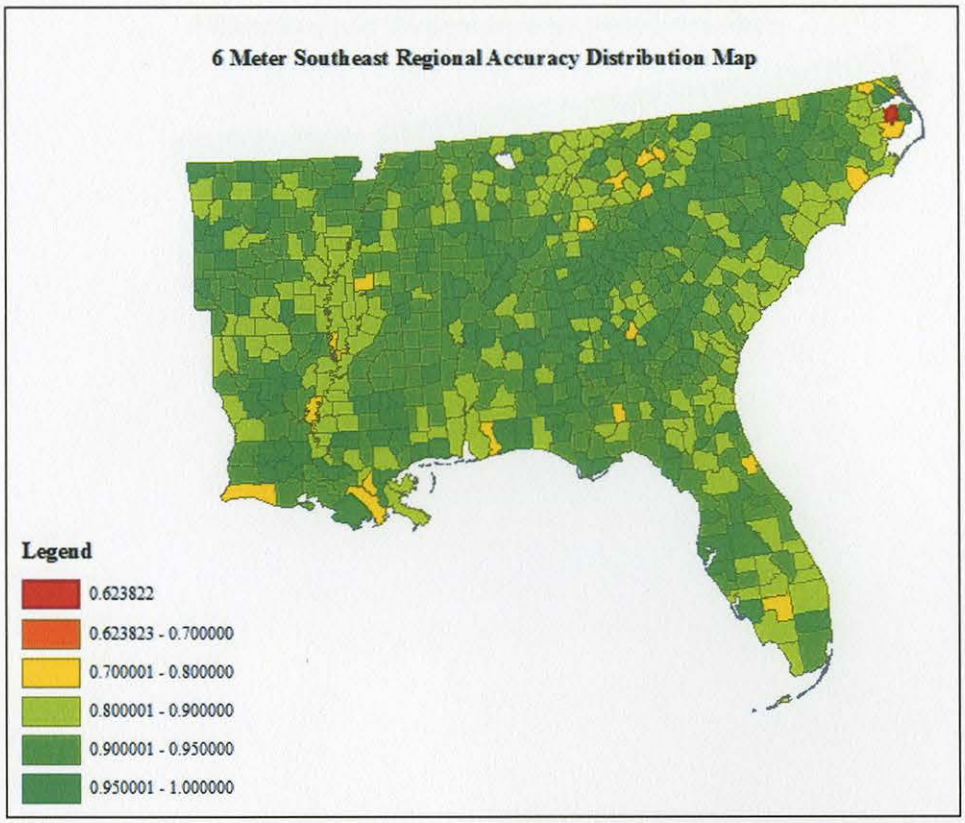

Figure 7. 6 meter buffer distance accuracy assessment map. This figure shows considerable improvement in typical OSM accuracy levels compared with the 4 meter buffer.

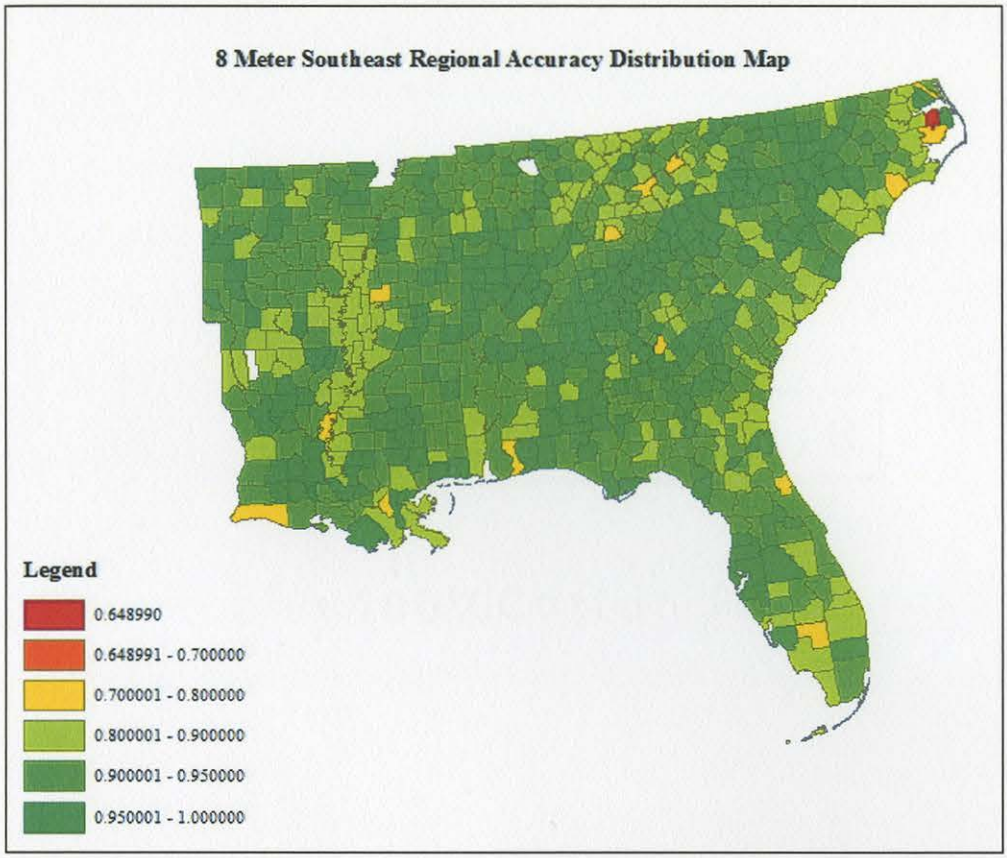

Figure 8. 8 meter buffer distance accuracy assessment map. Within an 8 meter buffer distance the functional accuracy of OSM data begins to average around 90% in the southeastern U.S.

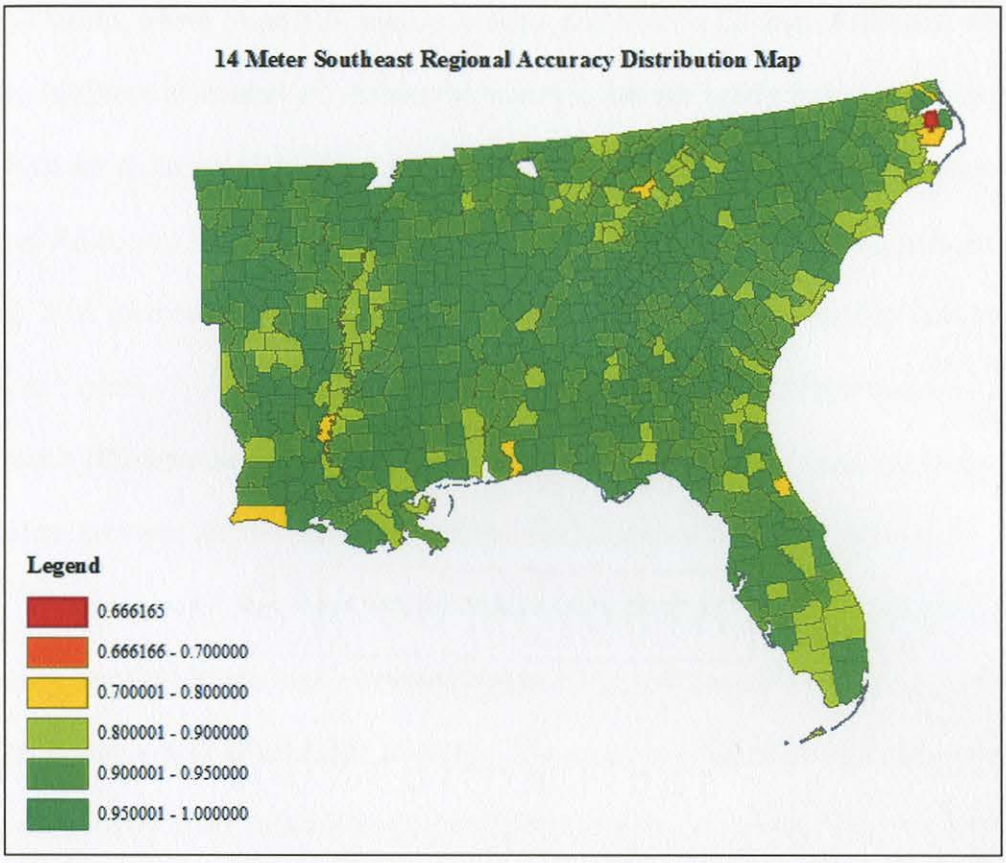

Figure 9. 14 meter buffer distance accuracy assessment map. This figure shows that the vast majority (approximately 94%) of OSM data lies within 14 meters of the matching Navteq features, indicating that OSM data is typically off by no more than the width of a four-lane road in many areas of the southeastern United States.

The comparison between the demographic and the OSM data accuracy levels revealed no statistical relationship showing that any of the tested demographics influence the positional accuracy of Open Street Map data. It was predicted that there would be a positive correlation between population density and Open Street Map Functional Accuracy levels due to the fact that Open Street Map data is entirely volunteer generated. However, this hypothesis has been proven incorrect through correlation analysis, as shown in the charts below population density has no statistical relationships to the functional accuracy level of OSM data. Further, looking at the maps in figures 6-9 it is clear that Saint Tammany Parish, Louisiana, a suburb area of New Orleans; Miami-Dade

county Florida, where Miami, Florida is located; and Mobile County, Alabama, where Mobile, Alabama is located are among the counties that are barely acceptable under the guidelines for accuracy assessment set out in the methodology. Other demographics tests included Education level (percent population that has a bachelor's degree), persons below poverty level (percent population), and age. The results for all demographic tests are shown in figures 10-18. Due to the data having a non-parametric distribution a Spearman's Rho correlation analysis was performed to determine if there is a true correlation between the two variables and the results were recorded in tables 6-8.

It was expected that there was a positive correlation between educational attainment and OSM functional accuracy levels. However it too showed no statistical correlation that would affect OSM accuracy. The same was found for percent population below the poverty level. It is believed that due to the inclusion of the U.S. Census bureau TIGER dataset (OpenStreetMap, 2011) as the basis for the OSM dataset for the United States that the OSM dataset is not significantly affected by socio-economic or political factors. Future research could explore this phenomenon to determine if the TIGER and OSM datasets remain functionally identical.

Effect of Population Density on the Accuracy of OSM Data

Figures 10-13 show the effect of population density on the functional accuracy of OSM data. The results show clearly that there is no statistical relationship between the two variables. This section continues by revealing the effects of educational levels, and income on the functional accuracy level. However, neither are revealing of a pattern. Several other fields were tested using this type of analysis but none were found to have a statistical relationship significant enough to act as a predictor of OSM accuracy.

Table 6

Spearman's Rho Correlation Analysis: Population Density vs Accuracy at Various Buffer Distances

Buffer Distance	Spearman's Rho (r_s)	Significance	Relationship
6 meter	-.068	.421	No statistically valid relationship
8 meter	-.057	.501	No statistically valid relationship
14 meter	-.069	.417	No statistically valid relationship

Note: Spearman's Rho analysis did not find a statistically valid relationship between population density and OSM data accuracy at any of the 10 buffered distances.

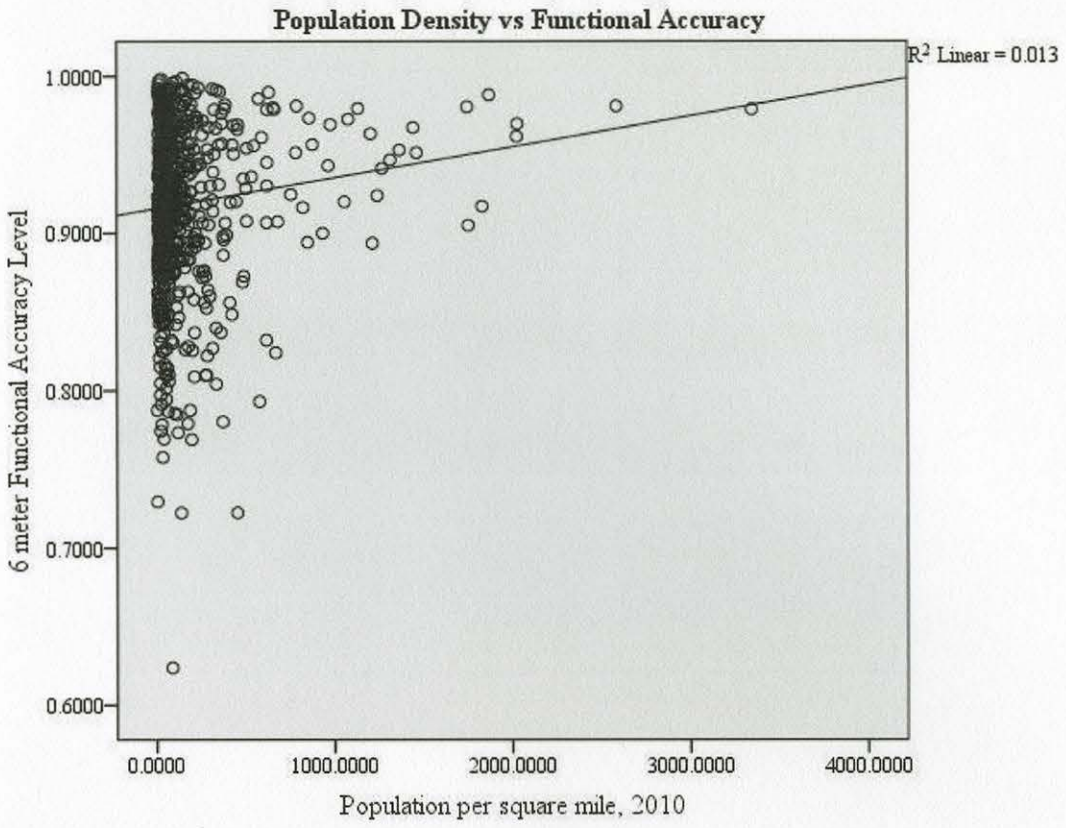

Figure 10. Population Density vs 6 Meter Buffered Distance Accuracy. Shows the relationship between the 6 meter buffered distance functional accuracy and the population per square mile.

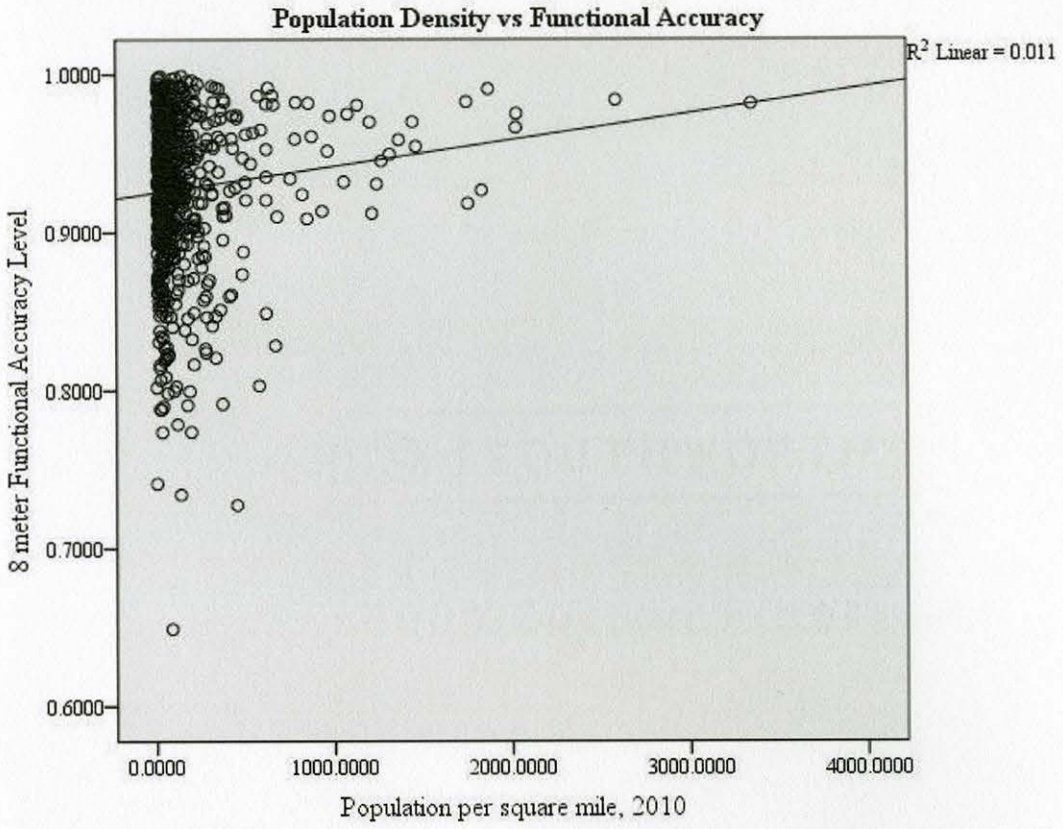

Figure 11. Population Density vs 8 Meter Buffered Distance Accuracy. Shows the relationship between the 8 meter buffered distance functional accuracy and the population per square mile.

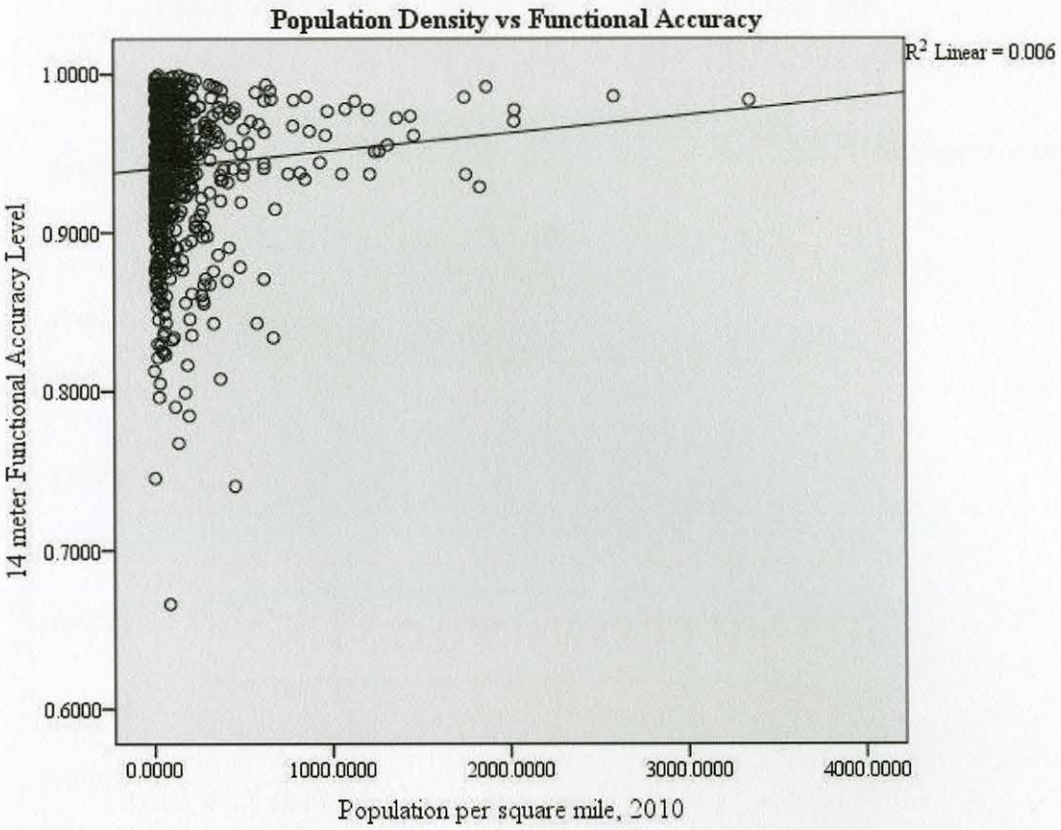

Figure 12. Population Density vs 14 Meter Buffered Distance Accuracy. Shows the relationship between the 14 meter buffered distance functional accuracy and the population per square mile.

Effect of Education Level on the Accuracy of OSM Data

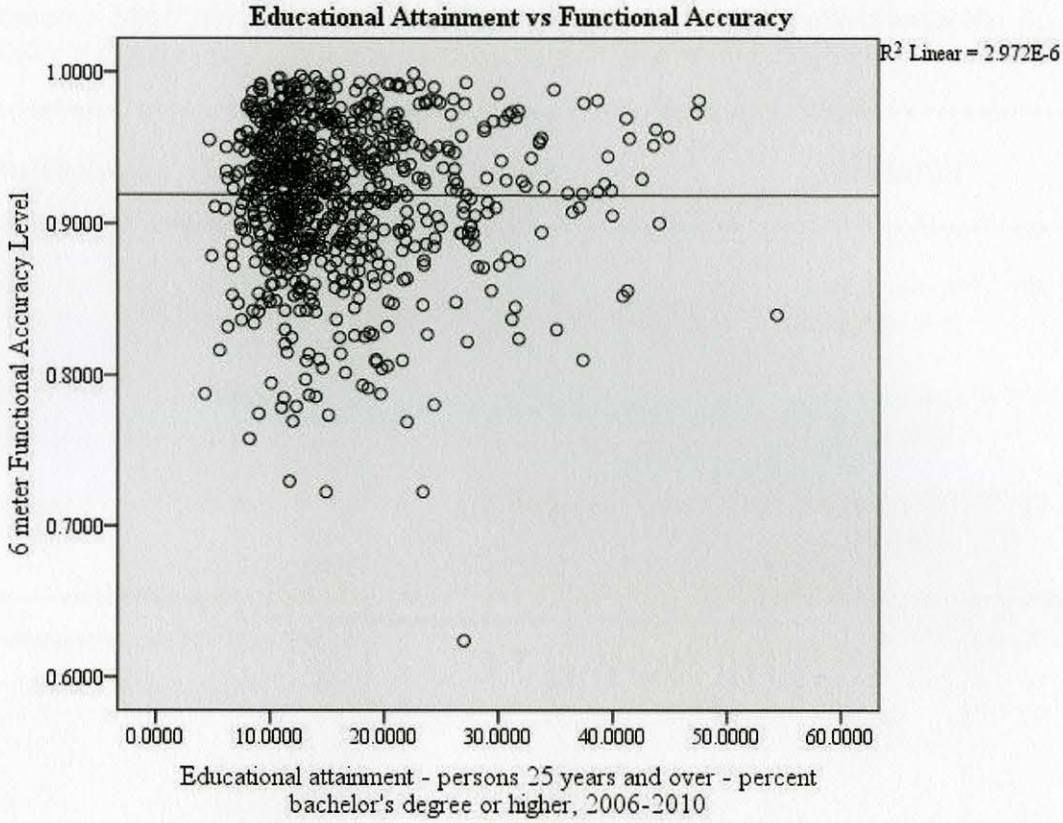

Figure 13. Education Level vs 6 Meter Buffered Distance Accuracy. Shows the relationship between the 6 meter buffered distance functional accuracy and the percentage of people with a bachelor's degree in the county.

Table 7

Spearman's Rho Correlation Analysis: Education Level (Percent Population with Bachelor's Degree) vs Accuracy at Various Buffer Distances

Buffer Distance	Spearman's Rho (r_s)	Significance	Relationship
6 meter	-.012	.885	No statistically valid relationship
8 meter	-.007	.934	No statistically valid relationship
14 meter	-.043	.618	No statistically valid relationship

Note: Spearman's Rho analysis did not find a statistically valid relationship between population density and OSM data accuracy at any of the 10 buffered distances.

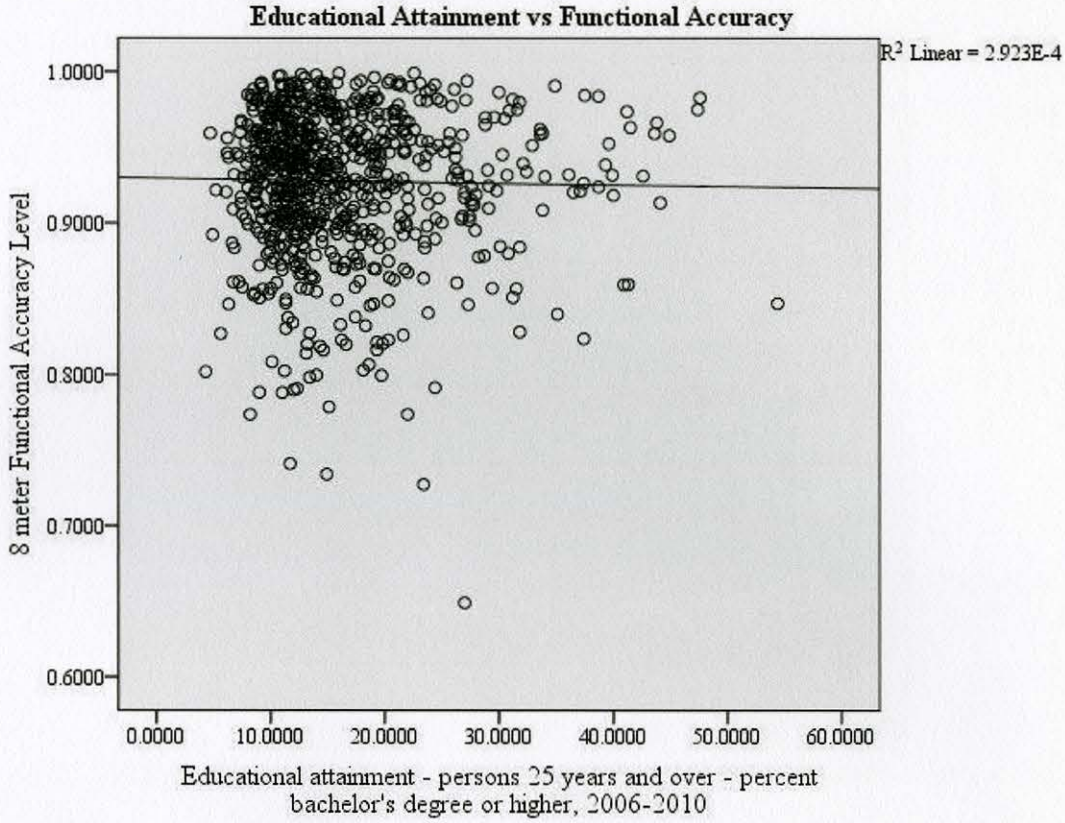

Figure 14. Education Level vs 8 Meter Buffered Distance Accuracy. Shows the relationship between the 8 meter buffered distance functional accuracy and the percentage of people with a bachelor's degree in the county.

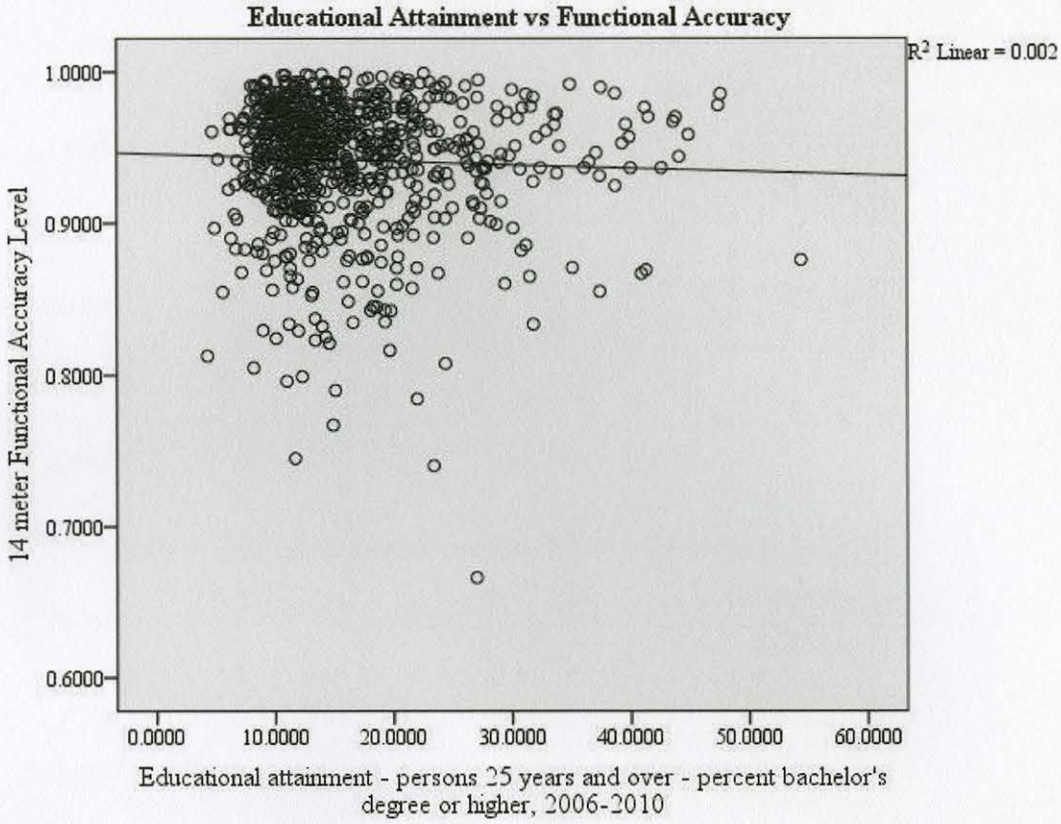

Figure 15. Education Level vs 14 Meter Buffered Distance Accuracy. Shows the relationship between the 14 meter buffered distance functional accuracy and the percentage of people with a bachelor's degree in the county.

Effect of Poverty on the Accuracy of OSM Data

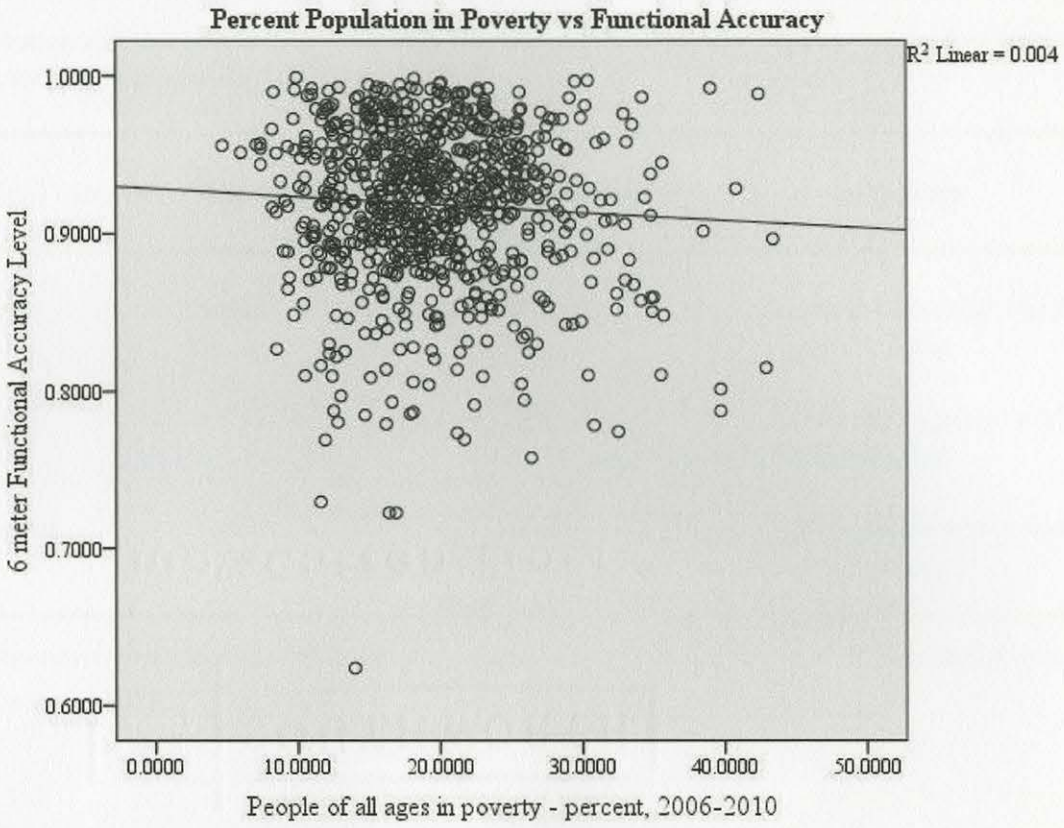

Figure 16. Percent Population in Poverty vs 6 Meter Buffered Distance Accuracy. Shows the relationship between the 6 meter buffered distance functional accuracy and the percentage of people who live in poverty.

Table 8

Spearman's Rho Correlation Analysis: Poverty Level (Percent Population in Poverty) vs Accuracy at Various Buffer Distances

Buffer Distance	Spearman's Rho (r_s)	Significance	Relationship
6 meter	-.068	.421	No statistically valid relationship
8 meter	-.057	.501	No statistically valid relationship
14 meter	-.069	.417	No statistically valid relationship

Note: Spearman's Rho analysis did not find a statistically valid relationship between population density and OSM data accuracy at any of the 10 buffered distances.

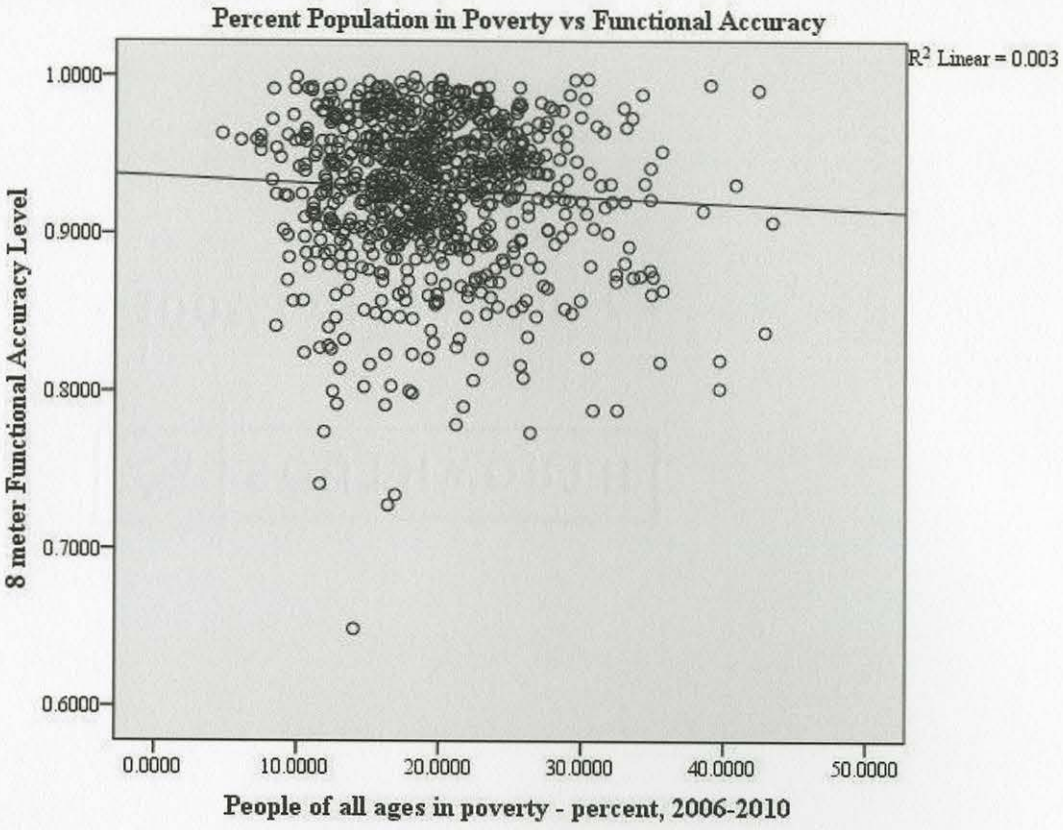

Figure 17. Percent Population in Poverty vs 8 Meter Buffered Distance Accuracy. Shows the relationship between the 8 meter buffered distance functional accuracy and the percentage of people who live in poverty.

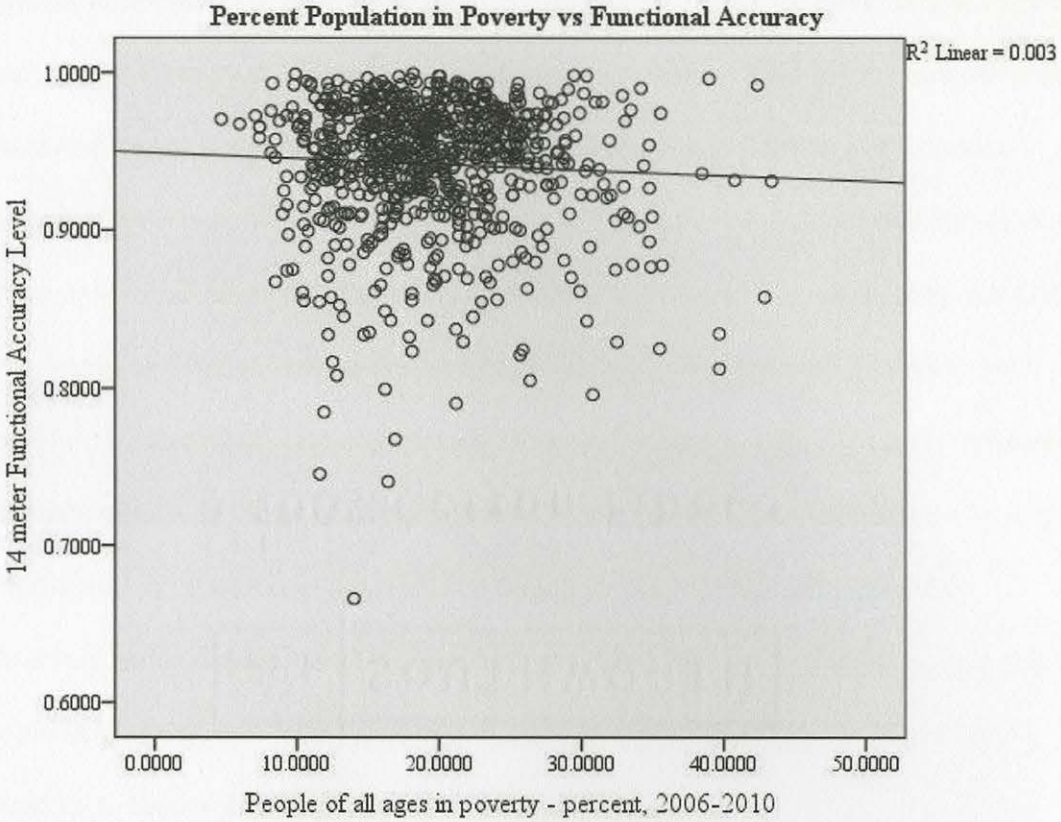

Figure 18. Percent Population in Poverty vs 14 Meter Buffered Distance Accuracy. Shows the relationship between the 14 meter buffered distance functional accuracy and the percentage of people who live in poverty.

Although OSM's positional accuracy has proven sufficient this does not constitute a completed quality analyses. The results of the manual analyses of the OSM and Navteq datasets for a limited number of counties based on the five ISO data quality metrics outlined in ISO 19113:2002 showed that while not as high quality as Navteq the acceptability of the quality of OSM data is dependent on the intended use of the product. In terms of positional accuracy OSM and Navteq datasets are close enough to be interchangeable based on this study, further OSM has more total road length in almost every county in the study indicating either (1) it has much higher geometric complexity, or (2) it has many road features that Navteq does not have. Comparing the datasets side

by side for a few select areas reveals that it is likely a little of both. For example Figures 19 and 20 show a section of Baldwin county Alabama where OSM has considerably more features and higher geometric complexity than Navteq. While Navteq also contains features that are not in the OSM dataset, it was common to find entire subdivisions and other complex road network areas in OSM that Navteq was missing indicating that OSM may be better in terms of unique feature count and temporal accuracy. However an inspection of the attribute tables shows that Navteq routinely collects more information about each feature. In terms of completeness the two datasets each have their advantages and the selection of which to use is highly dependent on whether a researcher or professional needs detailed road attribute information. The map in Figure 21 shows the counties in green where OSM had more total road length indicating that it has been updated to, in theory, better reflect real-world conditions.

Logical consistency is a metric that is strictly adhered to by both datasets, the required fields are always included, feature geometry and road type are the required components of any road network, and all other attributes are optional. As previously discussed one of the benefits of Open Street Map data is that it can be updated daily free of charge, to do the same with Navteq data there must be contracts in place to allow for updates to be received as frequently due to the commercial nature of the dataset. This ability to be updated readily indicates that for projects requiring a high degree of temporal accuracy OSM data may be a better choice than commercial options like Navteq. The existence of areas such as those shown in figures 19 and 20 that have significantly more data in the OSM dataset reveals that the accuracy levels are likely artificially deflated and that there are areas where OSM has a clear advantage due to the

frequency with which updates are issued. Likewise there are areas where Navteq has more roads but they are not as dramatically different. Thematic accuracy was not evaluated due to the fact that there is no available dataset to compare both OSM and Navteq data to that is of high, or at least known thematic accuracy levels.

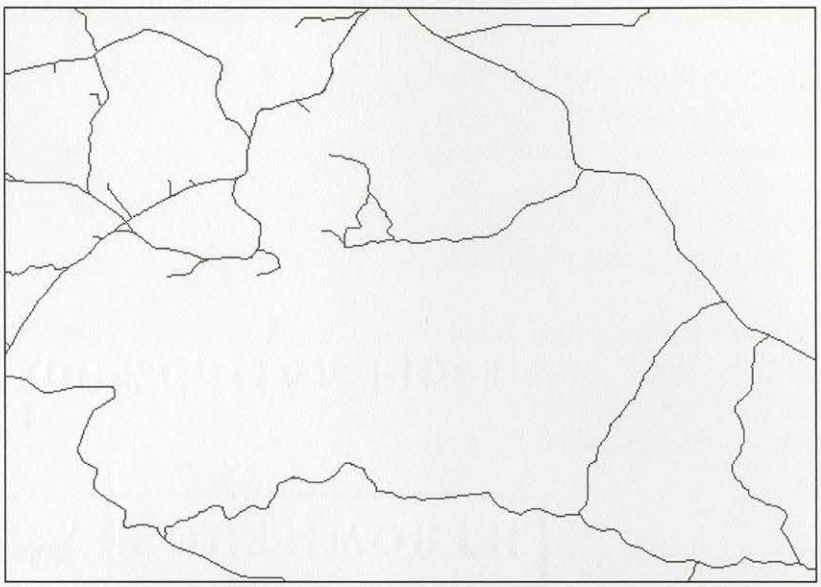

Figure 19. Baldwin County Alabama Navteq Dataset Roads .This figure shows the Navteq dataset for an area in northern Baldwin County, Alabama near Mobile, Alabama.

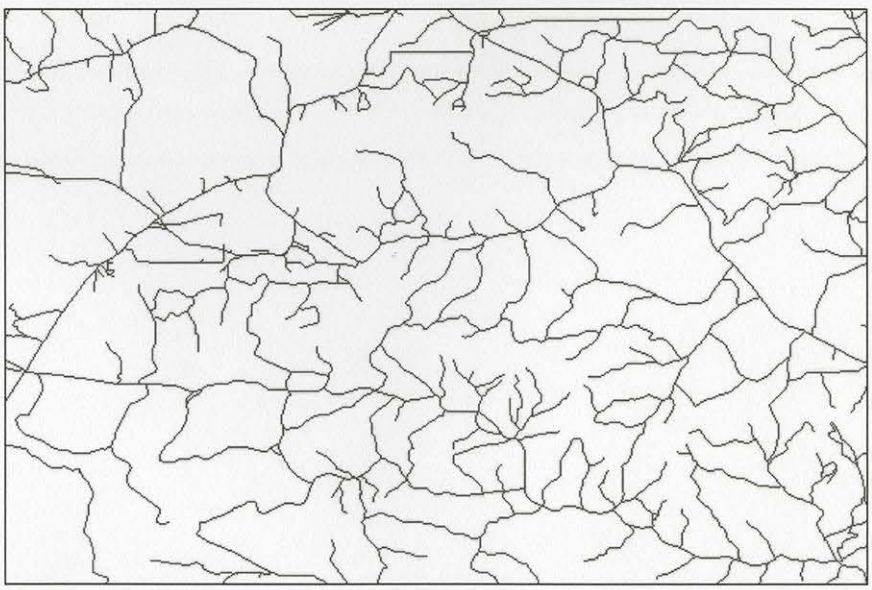

Figure 20. Baldwin County Alabama OSM Dataset Roads Shows the OSM dataset for the same area in northern Baldwin County, Alabama represented by Figure 19.

Table 9

Evaluation of ISO Quality Metrics.

ISO Quality Metric	OSM is better	Navteq is better	Notes
Completeness	X	X	Both are good for their own purposes, If a basic road map is all that is required for a project OSM is more complete and cheaper. However, Navteq is better for projects that are going to be used for routing and that require road attribute information.
Logical Consistency		X	Although both record a minimal set of data that is sufficient to draw road features, Navteq records its data in a much more consistent format and for more fields than OSM data.
Positional Accuracy	X	X	Per the accuracy analyses above at nearing 95% functional accuracy compared to Navteq within just 20 meters either is acceptable in terms of positional accuracy.
Thematic Accuracy			Not evaluated.

Note: This table is an evaluation of the datasets using ISO 19113:2002 criteria. Based on these results both datasets have their merits, the primary consideration for which to choose for a project would be the degree of attribute information that is required for the project.

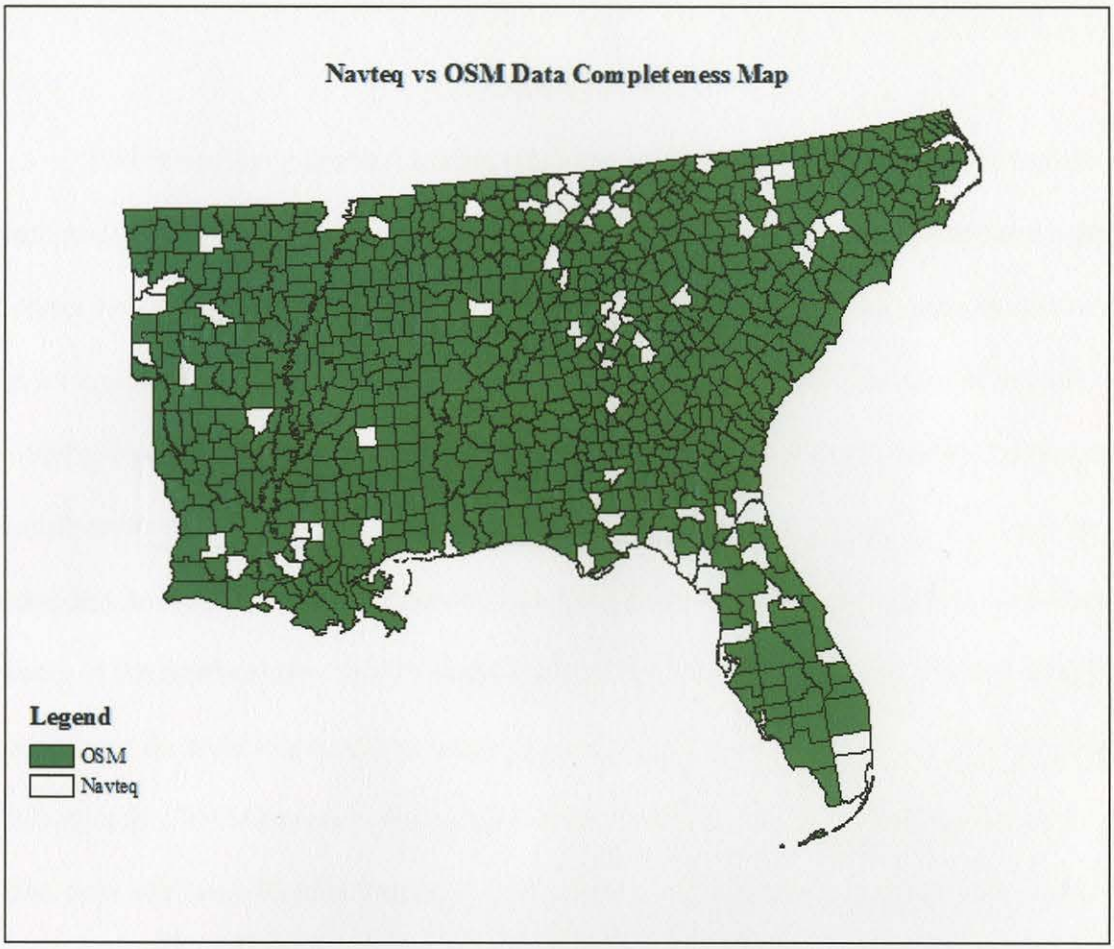

Figure 21. Navteq vs OSM Data Completeness Map. This map shows the areas in green where OSM has more total features in terms of total road lengths. It is clear that the OSM dataset contains more total feature length in the majority of counties.

CHAPTER V

DISCUSSION

This thesis has presented a complete view of the suitability of the OSM dataset compared with its commercial counterpart Navteq. Overall the findings have shown that the positional accuracy in the Southeastern United States appears to be quite similar to the accuracy of Navteq Data with accuracy levels ranging from 87.2% similar at a two meter buffer to 94.9% at a twenty meter buffer. These values indicate that having a strong starting point such as the TIGER dataset may increase the OSM accuracy. It should be noted that due to the fact that OSM datasets have more total road length in almost every county in the southeastern United States the accuracy values are naturally deflated and may be significantly higher if compared to ground truth data. The Hakley (2010) and the Ludwig et al. (2011) papers both indicated that in the United Kingdom and Germany OSM only achieved 80-85% data accuracy overall and had some areas that were much lower.

Despite having substantially higher positional accuracy than studies conducted in other areas it is clear that OSM data is not well suited for cases where detailed road attribute information is needed in the United States. For basic road-map applications OSM is perfectly well suited, and for cases where user annotation of features is necessary such as in disaster response situations it is arguably much better than commercial datasets. However, for critical need tasks such as emergency services routing, GPS based navigation, and similar tasks it will need improved attribute information. Based on the results of this study it can be concluded that Open Street Map is comparable in terms of positional accuracy to a defacto industry standard data product, Navteq, and it would be

of sufficient detail and quality to be of considerable use in many future GIS projects. It would be encouraged that future research be conducted to determine if OSM data is functionally identical to the TIGER dataset from the U. S. Census Bureau. An analyses of OSM compared to TIGER data would determine if the OSM dataset's positional accuracy and quality observed in this study is a factor of crowd sourcing geographic data or it if is simply a result of the fact that the OSM data was initially based on the TIGER dataset. If the OSM data is significantly different than the TIGER dataset it would indicate that the crowd sourcing or VGI effort has added value in increased temporal accuracy and completeness compared with traditional GIS data collection methods. The study performed for this thesis indicates that crowd-sourced/VGI data collection efforts can be comparable to commercial data products. The implications of this revelation can have significant impacts on the field of geography in three key ways. First, it is now possible to produce professional quality data products using amateur geographers making way for a potentially unlimited data acquisition for any topic of study. Second, the use of VGI data could force vendors to reduce costs, or increase quality of commercial products to differentiate themselves from the VGI equivalents of their products. Third, the available technologies that make VGI geographic data collection possible also encourages a spatial outlook in the clients that participate opening opportunities to bring more expertise and interest to the various sub-disciplines of geography. It is clear that as much as open source software has changed the computing world open data has the potential to change the scientific world by opening new opportunities through data availability.

APPENDIX A

DATA PREPARATION APPLICATION OUTPUT

Column	Data Contents
State	FIPS State Name
County	FIPS County Name
navteqFeatLen	Original Navteq Feature Length
originalFeatLength	Original OSM Feature Length
originalFeatCount	Original OSM Feature Count
OSMPercCoverage	Original OSM Percent Coverage
a2mBuff	Quantity of OSM data in meters within a 2 meter buffer
a2mFeatCount	OSM feature count within a 2 meter buffer
a2mPerc	Percent of original OSM data that is within a 2 meter buffer
a4mBuff	Quantity of OSM data in meters within a 4 meter buffer
a4mFeatCount	OSM feature count within a 4 meter buffer
a4mPerc	Percent of original OSM data that is within a 4 meter buffer
a6mBuff	Quantity of OSM data in meters within a 6 meter buffer
a6mFeatCount	OSM feature count within a 6 meter buffer
a6mPerc	Percent of original OSM data that is within a 6 meter buffer
a8mBuff	Quantity of OSM data in meters within a 8 meter buffer
a8mFeatCount	OSM feature count within a 8 meter buffer
a8mPerc	Percent of original OSM data that is within a 8 meter buffer
a10mBuff	Quantity of OSM data in meters within a 10 meter buffer
a10FeatCount	OSM feature count within a 10 meter buffer
a10mPerc	Percent of original OSM data that is within a 10 meter buffer
a12mBuff	Quantity of OSM data in meters within a 12 meter buffer
a12mFeatCount	OSM feature count within a 12 meter buffer
a12mPerc	Percent of original OSM data that is within a 12 meter buffer
a14mBuff	Quantity of OSM data in meters within a 14 meter buffer
a14mFeatCount	OSM feature count within a 14 meter buffer
a14mPerc	Percent of original OSM data that is within a 14 meter buffer
a16mBuff	Quantity of OSM data in meters within a 16 meter buffer
a16mFeatCount	OSM feature count within a 16 meter buffer
a16mPerc	Percent of original OSM data that is within a 16 meter buffer

a18mBuff	Quantity of OSM data in meters within a 18 meter buffer
a18mFeatCount	OSM feature count within a 18 meter buffer
a18mPerc	Percent of original OSM data that is within a 18 meter buffer
a20mBuff	Quantity of OSM data in meters within a 20 meter buffer
a20mFeatCount	OSM feature count within a 20 meter buffer
a20mPerc	Percent of original OSM data that is within a 20 meter buffer

APPENDIX B

U. S. CENSUS DEMOGRAPHIC FIELDS

Data Item	Item Description
STATECOU	FIPS State and County code
PST045211	Resident total population estimate (July 1) 2011
POP010210	Resident total population, 2010
POP050210	Resident total population, percent change - April 1, 2000 to April 1, 2010
POP010200	Resident population (April 1) 2000 (complete count)
AGE115210	Resident population under 5 years, percent, 2010
AGE275210	Resident population under 18 years, percent, 2010
AGE765210	Resident population 65 years and over, percent, 2010
SEX205210	Resident population: total females, percent, 2010
RHI105210	Resident population: White alone, percent, 2010
RHI205210	Resident population: Black alone, percent, 2010
RHI305210	Resident population: American Indian and Alaska Native alone, percent, 2010
RHI405210	Resident population: Asian alone, percent, 2010
RHI505210	Resident population: Native Hawaiian and Other Pacific Islander alone, percent, 2010
RHI605210	Resident population: Two or more races, percent, 2010
RHI705210	Resident population: Hispanic or Latino Origin, percent, 2010
RHI805210	Resident population: Not Hispanic, White alone, percent, 2010
POP715210	Population 1 year and over by residence - same house, one year ago, percent, 2006-2010
POP645210	Place of birth, foreign born, percent, 2006-2010
POP815210	Population 5 years and over, percent speaking language other than English at home, 2006-2010
EDU635210	Educational attainment - persons 25 years and over - percent high school graduate or higher, 2006-2010
EDU685210	Educational attainment - persons 25 years and over - percent bachelor's degree or higher, 2006-2010
VET605210	Veterans - total, 2006-2010
LFE305210	Average travel time to work for workers 16 years and over not working at home, 2006-2010
HSG030210	Housing unit, 2010
HSG445210	Owner-occupied housing units - percent of total occupied housing units, 2006-2010
HSG096210	Housing units by units in structure - multi-dwelling structure, percent, 2006-2010
HSG495210	Median value of specified owner-occupied housing units, 2006-2010

HSD410210	Households, 2006-2010
HSD310210	Average household size, 2006-2010
INC910210	Per capita income in the past 12 months (in 2010 inflation-adjusted dollars), 2006-2010
INC110210	Median household income, 2006-2010
PVY020210	People of all ages in poverty - percent, 2006-2010
BZA010209	Private nonfarm establishments, 2009
BZA110209	Private nonfarm employment for pay period including March 12, 2009
BZA115209	Private nonfarm employment for pay period including March 12, 2009, percent change, 2000-2009
NES010209	Nonemployer: total (NAICS 00) - establishments, 2009
SBO001207	Total number of firms, 2007
SBO315207	Total Black-owned firms, percent, 2007
SBO115207	Total American Indian- and Alaska Native-owned firms, percent, 2007
SBO215207	Total Asian-owned firms, percent, 2007
SBO515207	Total Native Hawaiian- and Other Pacific Islander-owned firms, percent, 2007
SBO415207	Total Hispanic-owned firms, percent, 2007
SBO015207	Total Women-owned firms, percent, 2007
MAN450207	Manufacturing: total (NAICS 31-33) - value of shipments, 2007
WTN220207	Wholesale trade: merchant wholesalers (NAICS 42) - sales of establishments with payroll, 2007
RTN130207	Retail trade: total (NAICS 44-45) - sales of establishments with payroll, 2007
RTN131207	Retail trade: total (NAICS 44-45) - sales of establishments with payroll per capita, 2007
AFN120207	Accommodation and Food Services: total (NAICS 72) - sales of establishments with payroll, 2007
BPS030210	New private housing units authorized by building permits - total, 2010 (20,000-place universe)
FED110209	Federal Government expenditure - total, FY 2009
LND110210	Land area in square miles, 2010
POP060210	Population per square mile, 2010

REFERENCES

Alts.net. 2008. Historical Notes about the Cost of Hard Drive Storage Space. *A Little Technology Shoppe.* Last accessed 1 April 2010. http://www.alts.net/ns1625/winchest.html

Elwood, S. 2009. Geographic Information Science: new geovisualization technologies -- emerging questions and linkages with GIScience research. *Progress in Human Geography*. 256-263.

ESRI. 2010. The Role of Volunteered Geographic Information in a Postmodern GIS World. *ArcUser.* Spring 2010 pp. 20-21.

European Commission. 2010. Haiti earthquake January 2010: damage assessment. *The European Commission Joint Research Centre*. Last accessed 19 April 2010. Last updated 17 March 2010. http://ec.europa.eu/dgs/jrc/index.cfm?id=5620

Federal Geographic Data Committee. 1998. Geospatial Positioning Accuracy Standards Part 3: National Standard for Spatial Data Accuracy. *United States Government*.

Feilner, M. 2009. OpenStreetMap Now in 3D. *Linux Magazine*. March 2009.

Garmin. 2012. North America - NAVTEQ Traffic. *Garmin*. Last Accessed 16 March, 2012. www.garmin.com/traffic/fm/navteq.html

Goodchild, M. F. 2007. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*.

Goodchild, M., & Hunter, G. 1997. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 299-306.

Google. 2011. Google Map Maker. Google.Last accessed 1 May 2011 www.google.com/mapmaker

Hakley, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environmental & Planning B: Planning & Design*, 682-703.

Hakley, M., & Weber, P. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 12-18.

Hall, B. G., Chipeniuk, R., Feick, R. D., Leahy, M. G., & Deparday, V. 2010. Community-based production of Geographic information using open source software and Web 2.0. *Geographical Information Science*, 761-781.

Kresse W., and Fadaie, K. 2004. *ISO Standards for Geographic Information*. Berlin: Springer-Verlag.

Linux Pro Magazine. 2010. Updated Map of Earthquake Region. *Linux Pro Magazine*, May 2010 p. 12.

Ludwig, I., Voss, A., & Krause-Traudes, M. 2011. A comparison of the street networks of Navteq and OSM in Germany. In: *Advancing Geoinformation Science for a Changing World* , pp. 65-84. Berlin: Springer.

Maron, M. 2010. Brain Off: Building Digital Technology for Our Planet. *Haiti OpenStreetMap Response*. Last accessed 12 February 2011. Last updated 14 January 2010. http://brainoff.com/weblog/2010/01/14/1518

Menga, R. 2007. Garmin and NAVTEQ Seal the Deal for the Long Haul. *PC Mech.* Last accessed 16 March 2012. Last updated 26 November 2007. http://www.pcmech.com/article/garmin-and-navteq-seal-the-deal-for-the-long-haul/

Microsoft. 2012. Bing Maps. *Microsoft Corporation*. Last accessed 14 March 2012 http://www.bing.com/maps/

National Geospatial Intelligence Agency. 2005. *Accessing Aeronautical Information from NGA*. U.S. Government.

National Oceanic and Atmospheric Administration. 2010. Environmental Response Management Application: Caribbean. *Environmental Response Management Application* Last accessed 20 April 2010. http://caribbean.erma.unh.edu/

Navteq. 2011. Map Reporter. *Navteq*. Last accessed 1 May 2011, http://mapreporter.navteq.com

Open Source Initiative. 2010. About the Open Source Initiative. *Open Source Initiative* Last accessed 25 April 2010. http://www.opensource.org/about

OpenStreetMap. 2010a. Garmin/GPS series. *OpenStreetMap*. Last accessed 20 April 2010. Last updated 18 April 2010. http://wiki.openstreetmap.org/wiki/Garmin/GPS_series

OpenStreetMap. 2010b. OpenStreetMap Map Making Overview. *OpenStreetMap* Last accessed 9 April 2010. Last updated 3 March 2010 http://wiki.openstreetmap.org/wiki/Map_Making_Overview

OpenStreetMap. 2010c.OpenStreetMap Quality Assurance. *OpenStreetMap.* Last accessed 9 April 2010. Last Updated 21 February 2010. http://wiki.openstreetmap.org/wiki/Quality_Assurance

OpenStreetMap. 2010d. OpenStreetMap Stats. *OpenStreetMap.* Last accessed 10 April 2010. Last updated 12 February 2010. http://wiki.openstreetmap.org/wiki/Stats

OpenStreetMap. 2010e. WikiProject Haiti/Press info. *OpenStreetMap.* Last accessed 9 February 2012. Last updated 24 January 2010. http://wiki.openstreetmap.org/wiki/WikiProject_Haiti/Press_info

OpenStreetMap. 2011. WikiProject United States/Data. *OpenStreetMap*. Last accessed 1 February 2012. Last Updated 15 December 2011. http://wiki.openstreetmap.org/wiki/WikiProject_United_States/Data

OpenStreetMap. 2012. Humanitarian OSM team. *OpenStreetMap.* Last accessed 19 February 2012. Last updated 10 February 2012. http://wiki.openstreetmap.org/wiki/Humanitarian_OSM_Team

Pentland, W. 2008. The World's Top Car-Owning Countries. *Forbes.* Last accessed 8 August, 2011. Last Updated 30 July 2008. http://www.forbes.com/2008/07/30/energy-europe-automobiles-biz-energy-cx_wp_0730cars.html

Perkins, C. 2007. Community Mapping. *The Cartographic Journal.*

Privat, L. 2011. Garmin Switches to NAVTEQ Maps in South Africa. *GPS Business News*. Last Accessed 14 March, 2012. Last Updated 18 April, 2011. http://www.gpsbusinessnews.com/Garmin-Switches-to-NAVTEQ-Maps-in-South-Africa_a2967.html

Science Commons. (2008). Principles for Open Science. *Science Commons.* Last accessed 5 April 2010. Last updated 2008. http://sciencecommons.org/resources/readingroom/principles-for-open-science/

Science Daily. 2007. Mapmaking For The Masses: User-Generated Content Can Profoundly Impact Geographic Information Systems. *Science Daily*. December 2007.

State of Florida. 2012. Plans Preparation Manual. Tallahassee, Florida

State of Minnesota. 1999. Positional Accuracy Handbook. *Minnesota Planning Land Management Information Center.* St. Paul, Minnesota.

United States Department of Agriculture. 2004. Economic Research Service. *Measuring Rural-Urban Continuum Codes.* Last accessed 5 May 2011. Last updated 28 April 2004. http://www.ers.usda.gov/briefing/rurality/

Unitar. 2012. United Nations Institute for Training and Research. *United Nations*. Last accessed 1 February 2012. Last updated February 2012. www.unitar.org

Van Niel, T. G., and McVicar, T. R. 2002. Experimental evaluation of positional accuracy estimates from a linear network using point and line based testing methods. *International Journal of Geographical Information Science*. 455-473.

Wood, J. 2005. 'How green is my valley?' Desktop Geographic Information Systems as a community-based participatory mapping tool. *Area.*

Zandbergen, P. A. 2008. Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy. *Transactions in GIS*. 12(1):103-130.

Zielstra, D. and Zipf, A. 2010. Quantitative Studies on the Data Quality of OpenStreetMap in Germany. *Proceedings of the GIScience 2010 Conference.* 14 - 17 September 2010.